THE MOLECULAR ETIOLOGY OF ATYPICAL CYSTIC FIBROSIS

&

PHENOTYPIC HETEROGENEITY IN CLASSIC CYSTIC FIBROSIS

by
Briana Vecchio-Pagán

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
May, 2016

Abstract

This thesis highlights two studies which seek to link phenotypic presentations in cystic fibrosis (CF) and CF-like disease with underlying causative variation. Cystic fibrosis is a rare Mendelian disorder with a primary phenotype of highly viscous secretions in the lung, pancreas, and intestines. The current life expectancy for CF patient is ~50 years or age, and the primary cause of mortality is progressive lung disease. Individuals with this disorder carry two disease causing alleles in the cystic fibrosis transmembrane conductance regulatory gene (*CFTR*). However, there are individuals which present with a similar but mild CF-like disease who completely lack causative mutations at the *CFTR* locus. In **Chapter 2**, I report the discovery and functional analysis of mutations within a candidate causative locus, *CA12*, in these 'atypical' CF patients. Our study details three novel variants in *CA12* or carbonic anhydrase XII which are inherited in a recessive fashion in two unrelated pedigrees. We show that these variants, as well as previously reported variants, are complete loss of function through extensive localization and functional studies. Given *CA12*'s role in bicarbonate buffering, this study sheds new light on the importance of bicarbonate in CF disease.

There are over 2000 known variants within *CFTR*. However, the most common causative allele is deletion of phenylalanine at amino acid position 508, often referred to as F508del. Nearly 50% of all CF patients are homozygous for this allele, yet these patients exhibit a range of disease severity or phenotypic heterogeneity. In **Chapter 3** of this work, I present a study which sought to discover rare and previously untyped common variation within and surrounding *CFTR* that could modify disease severity in the F508del homozygous CF population. My findings suggest that a burden of variants which associate with CF traits exists at key loci throughout the *CFTR* genomic region. These loci include previously identified CTCF binding sites and other regulatory regions of interest. This study demonstrates the utility of re-seqeuncing a disease causing locus, and helps further our understanding of intragenic genetic modifiers.

Advisor & First Reader: Garry R. Cutting, M.D.

Professor, Department of Pediatrics

McKusick-Nathans Institute for Genetic Medicine

Johns Hopkins University School of Medicine

Second Reader: Loyal Goff, Ph.D.

Assistant Professor of Neuroscience

Johns Hopkins University School of Medicine

Thesis Committee:

Andrew McCallion, Ph.D. (Chair)

Dr. Scott Blackman, M.D./Ph.D.

Dr. Steven Salzberg, Ph.D.

Acknowledgements

I'd first and foremost like to acknowledge my thesis mentor, Dr. Garry Cutting. Dr. Cutting provided me with the opportunity to study computational biology and utilize my burgeoning skills on exceptional datasets. He has provided unwavering support during my graduate studies, and his confidence in my abilities often encouraged me as I struggled with challenging assays. He has further taught me the importance of growing research through collaboration, a skill which certainly will be valuable in my future career.

Next, I would like to thank Dr. Scott Blackman for his excellent mentorship. Scott has taken great time and care introducing me to datasets, walking me through various assays, and evaluating my results. He has also gone beyond the science, providing me with speaking opportunities and teaching valuable networking skills. Scott's enthusiasm for science and ability to put a positive spin on any negative result is addicting -- it is a trait I can only hope exhibits lateral gene transfer.

Melissa Lee started in the Cutting lab with me in 2011, and has since become a great friend. She is an incredibly thoughtful and highly motivating coworker. Melissa has encouraged me to have a strong voice and always welcomes my opinions and dark humor. She has pulled me through times when assays were not working, and together we have learned valuable lessons about graduate research. Finally, Melissa is truly an innovative scientist, and I have no doubt she is headed toward a prosperous career in this field.

I'd also like to thank many of the other Cutting laboratory members and alumni. This includes Laura Gottschalk, a very hard working scientist, but also caring mother of two. Her ability to balance her graduate studies with home life was inspiring to me, and her sympathy and advice with regards to both are unparalleled. My research would not have been possible without Karen Raraigh, who tirelessly recruits patients, coordinates blood draws, delves into the phenotypes, and basically holds everything together on a daily basis. I will certainly miss discussing new findings with her, among other non-research related genetic endeavors.

I'd also like to thank Dr. Neeraj Sharma, whose depth of knowledge and critical thinking skills have helped me work through many scientific challenges. He is an invaluable asset to our laboratory.

Finally, I'd like to thank the remainder of my colleagues including Jeenah Park, Arianna Franca, Mike Collaco, Patrick Sosnay, Allison McCague, Ted Han, Anh-Thu Lam, Matthew Pellicore, Taylor Evans, Emily Davis, and Yasmine Akhtar. These individuals have spent many hours carefully pipetting samples for the research presented in this work, or have contributed their time and knowledge in too many ways to be endeavored here.

I thank the cellular and molecular medicine graduate training program (CMM) for always providing guidance. Colleen Graham and Leslie Lichter have provided excellent mentorship and were always ready to listen when I walked through the door. I'd like to acknowledge Tricia Cornwall, who diligently manages to keep the budgets balanced and find time in Garry's busy schedule for meetings. She is another invaluable resource that makes this research possible.

I'd like to thank my thesis committee members for their valuable guidance throughout my studies. I thank Dr. Loyal Goff for his thoughtful review of this manuscript. I'd also like to recognize Dr. Joginder Nath, my first genetics professor. His love for this field and excitement about its future propelled me into this career. I hope that he continues to share his passion with future generations.

My parents and sister have provided an unwavering support network for me throughout my education. They have made many large and small sacrifices such that I could pursue my ambitions. Most importantly, they recognized my love for science at a young age and ensured that I had the best possible teachers and other influences through-out my childhood. I am very proud to call myself a part of their family.

Finally, I'd like to thank my husband, Vincent Pagán, for sharing in my love of scientific research. Coming home to someone who understands the ins and outs of this career path, and is able to provide both sympathy and advice has made my journey much more enjoyable. I hope that we will always remember this special time together when we grew not only in our studies, but also our family. To Felix: My passion for science is only secondary to my love for you. It is my sincere wish that you will one day have a career doing what you love.

It is with my deepest gratitude and utmost respect that I dedicate this dissertation to my parents, who instilled in me the belief I could achieve anything in life. A career in genetics was my dream, and because of you, it begins with these pages.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

The Cystic Fibrosis Phenotype

Cystic fibrosis, or CF, is a disease affecting ~70,000 individuals worldwide, and nearly 30,000 individuals in the US (*1*). CF is a rare recessive Mendelian disorder, with a carrier frequency ranging from 1/30 to 1/17 in Caucasians of European descent, the primary affected population (*2*). Individuals with CF develop unusually viscous or otherwise altered secretions in several tissues including the lungs, pancreas, sweat duct, liver, intestines, and male reproductive tracts. These manifestations are further detailed in the following literature review, a natural history of CF.

In 1606, a Professor of Medicine at Henares in Spain noted that one's fingers would taste salty after rubbing the forehead of a 'bewitched' child (*3*). This is one of many similar early accounts of the later understood medical condition cystic fibrosis. This was a keen observation as the chloride concentration in sweat is one of the primary assays currently used to diagnose patients with CF. Healthy individuals typically have sweat chloride level measurements less than 40mM, whereas individuals with classic CF have measurements exceeding 60mM (*4*). Still, some individuals have intermediate values (40-60mM), and further testing is needed for diagnosis (see below: Atypical Cystic Fibrosis and **Chapter 2**) (*5*). While salty sweat was perhaps the first trait observed in CF, it was not fully appreciated until centuries later – likely because the primary cause of morbidity in these patients involved other tissues.

Pancreatic and intestinal manifestations of this disease had been reported as early as 1888, and were often mistaken as unusual cases of celiac disease (*6*). The term cystic fibrosis (CF) was originally developed in 1938 by Dorothy Anderson, and was first

introduced as “fibrocystic disease of the pancreas” (*7*). She observed patients with CF that had unusually thick secretions in the pancreas. These viscous secretions lead to pancreatic insufficiency in 85% of CF patients (*8*). The inability to digest and absorb nutrients was the primary cause of mortality in CF until the introduction of pancreatic enzymes in the late 1970s (*9*). A decrease in pancreatic function is evident at birth in many CF patients, and serum immunoreactive trypsinogen (a pancreatic enzyme precursor) levels are used today as a pre-natal screen for the disease (*10*).

Additional mortality in some CF patients was due to meconium ileus (MI). This is an intestinal blockage most common in CF patients, but present in the general population. It was originally described by Karl Freiherr von Rokitansky in 1838 following the autopsy of an infant with “yellow jelly-like meconium…sticking to the outside of the bowel” (*3*). Mortality of CF patients with MI was nearly 70% by age one in 1940 (*11*). Advances in detection and surgical correction of this blockage have greatly reduced mortality in patients with MI to that of the general CF population.

Once CF patients were living beyond early childhood (primarily due to pancreatic enzymes correcting malabsorption), an additional trait became the major cause of mortality: lung disease. Perhaps the first account of a CF patient with lung disease occurred in 1934 by a Dutch physician, GJ Huet (*3*). In addition to pancreatic insufficiency, the patient he reported had chronic cough beginning at age six, and died of bronchopneumonia at age 10. Lung disease was later described more formally in 1938 in Dr. Dorothy Anderson’s review of her own patients and literature case reports, “For weeks or months it remained as a chronic cough, gradually increasing in severity, with development of purulent bronchitis, bronchiectasis with abscesses, lobular pneumonia or

any combination of these conditions" (*12*). Lung disease due to recurrent infections followed by inflammation leads to eventual pulmonary failure and mortality in CF patients today.

In 1944, a great breakthrough in the treatment of lung infections was made available to CF patients, Penicillin (*13*). When this and other antibiotics were introduced to the treatment regimen of CF patients, clinical course improved significantly (*14*). This and other modern treatments such as physical therapy, inhaled steroids, and mucolytics have increased the average life expectancy for patients with CF to nearly 50 years of age (for children born today and projecting the same rate of progress in therapeutics) (*15*).

Molecular Etiology of Cystic Fibrosis

While the clinical course of CF was well appreciated in the early 20th century, understanding the underlying defect leading to symptoms would take much longer. The unusual pathology of the lung epithelium in CF patients was first described in 1923 as "A striking finding in the lungs was the epithelial lining of the bronchioles, which in many instances was definitely of a stratified squamous type and showed desquamation of the horny superficial portions". The same manuscript details unusual epithelial surfaces in other CF related tissues like the pancreas (*16*). These findings were eventually thought to be caused by the thick mucus and secretions in the same tissues. In the 1930s, a connection was established between abnormal lung epithelia and frequent infections (*17*). The observation linking lung disease to unusual epithelial morphology was first made for Vitamin A deficiency, part of the differential diagnosis of CF at the time. It was still

nearly a decade before additional insight into the molecular etiology of this disease was gained.

In 1951, Dorothy Anderson and Walter Kessler noticed an unusually high frequency of cystic fibrosis patients being admitted for dehydration during a heat wave (*18*). At this time, they speculated that perhaps it was a deficiency of the sweat gland in these patients, who (as noted above) had very salty sweat. The abnormally high sweat chloride concentrations in these patients were first quantified in 1953(*19*). Sweat testing as a diagnostic tool was later standardized by Gibson and Cooke (*20*). This was the first insight into a possible cause of the disorder; however, it was unappreciated for nearly another two decades as scientists pursued other hypotheses (*21*). These included the possibility of diffusible "factors" which may somehow alter CF patient lung secretions.

Significant enhancements were made in understanding the basic defect of CF in the 1980s. The first report of a possible defect in epithelial ion transport was on a poster presented at the 8th international congress on CF (*22*), "the existence of an extraordinarily active sodium absorption could explain some of the clinical findings of CF e.g. hyperviscous mucus as being caused by excessive absorption of NaCl and water". One of the pivotal papers which significantly increased our understanding of CF pathology was that of Dr. Michael Knowles and his team at the University of North Carolina. In 1981, he reported that the potential difference across nasal epithelia in CF patients was aberrant (*23*). Specifically, he was able to show that transport of both chloride and sodium was altered through use of specific inhibitors, and went on to hypothesize that this could be the primary defect in CF patients. Because these findings were observed in CF patients shortly after birth, it ruled out previous theories of diffusible factors causing

disease, and suggested a strong genetic component. Just a few years later, Paul Quinton (a renown CF researcher and CF patient) published another landmark paper detailing specifically a block of ion transport in the sweat duct (*24*). Later, Dr. Ray Frizzell confirmed the presence of a chloride channel which was defective in CF epithelial cells (*25*). We now know that defects in this channel, the cystic fibrosis transmembrane conductance regulator (*CFTR*), are the cause of CF.

In recent years, our understanding of the molecular etiology of CF has been further refined. CFTR is a large ATP-gated anion channel which is found within the apical membrane of secretory epithelial cells in the lung, liver, pancreas, intestines, skin, and reproductive tract. While its structure is similar to an ABC transporter, it functions as a channel with only passive transport. It is capable of conducting chloride, bicarbonate, and thiocyanate down their concentration gradient (*26*).

In the lungs and pancreas, CFTR functions to regulate chloride flow from the hypertonic intracellular fluid into the hypotonic secretions above epithelial cells. The movement of chloride occurs concurrently with the movement of sodium and water along their respective gradients, allowing for hydration of the epithelial surface. In the lung, the height of the airway surface liquid (ASL) and composition of the mucus layer allows cilia at this surface to beat correctly, transporting mucus (and the constantly inhaled bacteria and allergens contained there-in) out of the lung and into the digestive tract. When CFTR is not present at the membrane, or not functional, chloride transport is disrupted. This leads to dehydration of the mucus layer, inhibiting their ability of cilia to beat correctly and to clear particles trapped in mucous. Because of this reduced clearance, there is a build-up of pathogens in the lungs (i.e. *Pseudomonas aeruginosa*),

ultimately leading to severe lung disease. CFTR functions slightly differently in the sweat gland, where its primary role is resorption of chloride ions from the sweat(*24*). In the sweat duct, defective CFTR results in a hypertonic solution as chloride is trapped outside of the epithelial layer. This results in the salty sweat so characteristic of these patients, which continues to be used as a key diagnostic marker today.

Discovery of the *CFTR* gene

Much of our understanding of the role of the CFTR protein stems from genetic studies of CF. In 1985, family based genetic linkage studies were used to isolate the "CF gene" to the long arm of chromosome 7 (*27-30*). In 1988, the gene had been further localized to 7q21-31, and haplotypes suggested there was one predominant mutation in Europeans (*31*).

Finally, in 1989, several groups published the location and cloning of the cystic fibrosis gene, which they dubbed the cystic fibrosis transmembrane conductance regulator (*32-34*). With the cloning and characterization of *CFTR*, significant insight was gained as to the structure and function of the protein. It was also confirmed that the most common variant was a deletion of phenylalanine (3bp) within the first nucleotide binding domain (*33*). This variant was present on a specific haplotype background in Europeans, and was primarily observed in pancreatic insufficient patients. The remaining haplotypes of this gene were of low frequency, and subsets of them were more likely to occur in pancreatic sufficient patients (*32*). This suggested there were many other disease-causing mutations within this gene yet to be discovered.

The Genotype - Phenotype Correlation in CF

We now know that over 2000 variants have been observed within *CFTR*, and some of these are known to be disease causing (*35*). The severity of disease in each patient is primarily determined by which CF-causing variants they have inherited (*36*). CF variants can be classified into several groups by their expected functional impact on the cystic fibrosis transmembrane conductance regulator, or *CFTR* (*37*).

The most common mutation, F508del, is present (on either 1 chromosome or both) in nearly 70% of CF patients, where it results in the improper folding and eventual degradation of the final protein product (*38*). Phenotypes of patients carrying this allele reflect the damage present at the molecular level. Patients with two copies of F508del (~50% of the CF patient population) have an average sweat chloride of 102mM, 98% are pancreatic insufficient, and 60% develop a *P. aeruginosa* infection (39).

Other disease causing variants result in less severe CF phenotypes. For example, in patients carrying one copy of F508del and one copy of S1251N, a milder phenotype is observed. Their average sweat chloride is 90 mM, only 78% are pancreatic insufficient, and 50% contract *P. aeruginosa* (39). It is not shocking that these patients have a milder phenotype given the mutation, S1251N, results in a stable protein at the epithelial surface with partial chloride conductance. Because S1251N CFTR has the primary defect of chloride conductance, it is often referred to as a "gating" mutation (40).

However, the molecular defect in CFTR (be it inability to conduct chloride, lack of protein at the cell surface, or various other defects) does not always match with the observed phenotype (i.e. mild mutations can result in severe phenotypes and vice versa)

(*35*). In these scenarios, we may hypothesize that our initial assumption about the molecular defect resulting from a given mutation is invalid. For example, a missense mutation near the end of an exon may result in mis-splicing of the forming mRNA transcript. This could result in degradation of the erroneous transcript, no protein product, and thus a severe phenotype. Given the large number of variants in the gene, we are just now beginning to investigate and understand the functional defect of many *CFTR* mutations. This is an active area of research in our own and others' laboratories.

Phenotypic Heterogeneity in Cystic Fibrosis

There is extensive phenotypic heterogeneity even across CF patients carrying the same mutations in *CFTR*. This phenomenon was initially noticed within pedigrees containing multiple affected individuals (*41, 42*). In these families, some siblings would present with more or less severe lung disease or pancreatic function. Given that siblings share many environmental factors, it is possible that the portion of genetic variation they do not share could alter their disease course. This idea was systematically explored through hereditary studies of twins and siblings.

Initial twin and sibling based studies of lung disease heritability in cystic fibrosis were somewhat underpowered, and did not find a significant genetic contribution to the variability of this trait (*43*). These studies speculated that shared factors such as genetics and environment likely did not contribute to lung disease severity (*44*).

In a paper by Burkhard Tümmler in 2001, the chloride permeability of intestinal tissue as well as respiratory epithelia was studied in a cohort of F508del homozygous twins and siblings. In this cohort, approximately 30% of patients had residual chloride

transport in their pulmonary tissues, and this correlated with milder disease. Similar findings were observed in the intestinal tissues (*45*). The authors speculated that these patients must have some functional CFTR present at the cell membrane. Additionally, they noted that the chloride transport profiles observed were more concordant in monozygotic twins than in dizygotic twins and siblings. Greater concordance indicated that genetic variation beyond the primary *CFTR* mutation (F508del in this case) was contributing to chloride secretion and disease severity.

In 2005, Garry Cutting and colleagues published a paper detailing 526 patients from the Johns Hopkins Twin and Sibling Study (TSS) (*46*). In this manuscript, they found that cross sectional and 5 year longitudinal CF specific lung function percentiles were significantly more concordant for the monozygotic twin cohort. These findings now strongly suggested a role for shared genetic factors, beyond the *CFTR* locus, contributing to lung disease severity.

The relative contributions of environmental, genetic, and stochastic factors have now been estimated for lung disease (*47*), as well as various other CF traits (*37*). For example, variation in the risk of developing intestinal obstruction is nearly completely attributable to genetic factors (*48*). Other traits such as *P. aeruginosa* rates have a larger environmental component (*49*). Variation in sweat chloride levels has not been formally studied for relative contribution of genetic or environmental components. However, work in the Cutting laboratory by J. Michael Collaco suggests that nearly 50% of the variability in this trait is derived from CFTR genotypes, with another 25% likely due to other genes, and a final 25% due to environmental or stochastic factors. These studies firmly established a role for genetic modifiers of a Mendelian disease.

Genetic Modifiers of Cystic Fibrosis Traits

Various loci were first suspected to modify CF traits. These were primarily chosen based on a logical connection to the phenotype. For example, the alpha1-antitrypsin gene was known to have low functioning variant forms in European Caucasians. Given that deficiency of this enzyme results in poor lung function in non-CF populations, it was a great candidate modifier of CF lung function. (*50*). While this candidate was initially promising in small cohorts, later studies could not replicate an association between variants in alpha1-antitrypsin and lung disease severity (*51*). For many years, choosing candidate loci based on a known function was the only method available to pursue modifier genes in CF (*52-54*). In 2000, a consortium was formed in the CF genetics community in order to more systematically assay modifier loci.

Advances in high throughput methods of genotyping patients led to the development of genome wide association studies (GWAS) in the early 2000s (*55*). These assays were developed to map genes which are associated with either common complex phenotypes (e.g. heart disease), or which modify Mendelian disease phenotypes. And unlike previous genetic association studies to date, the sequencing of the human genome and use of single nucleotide polymorphism (SNP) genotyping arrays allowed for a genome-wide search. The CF research community took notice of this advancement, and the first GWAS for lung disease severity in CF patients was published in 2011 (*56*). This paper by the international genetic modifier consortium detailed two key loci associated with lung disease severity: 11p13 and 20q13.2. A subsequent GWAS in a larger CF cohort later replicated the 11p13 locus, and introduced 5 additional loci which may modulate this trait (*57*). GWAS have been performed for other CF traits such as CF-

related diabetes and meconium ileus (*58, 59*). However, this type of study has never been performed for sweat chloride levels – presumably because the degree of variance in this trait attributable to loci beyond *CFTR* is not well understood (see above reference to ongoing work by JM Collaco).

The Search for Missing Heritability

A phenomenon that continues to perplex researchers today is the "missing heritability" present in GWAS (*60*). Through twin and sibling studies, one can estimate the proportion of variance in a given trait due to shared inherited genetics; this is often termed 'heritability'. Similarly in GWAS, the total percentage of variance explained by each key finding can be calculated. Missing heritability arises when the expected proportion of variance attributable to a trait through previous studies is significantly larger than the variance explained by GWAS. This discrepancy may be due to a variety of factors including genetic epistasis (interactions of multiple genes) (*61*), linkage masking (*62*), dominant heritability models (*63*), and rare variation (*64*). There is also missing heritability for lung disease severity in cystic fibrosis. The most recent GWAS identified 5 loci, each with beta values less than 5% for the forced expiratory volume (FEV1) percent predicted measure of CF lung function(*57*). The analyses in this manuscript indicate that there is still substantial missing heritability for this CF trait.

In an attempt to find additional genetic factors which may be contributing to lung disease severity in CF patients, **Chapter 3** of this manuscript details a novel re-sequencing study of the *CFTR* locus. In this study, we sought to determine whether

previously untyped variants, within and surrounding the causative gene, could lead to phenotypic heterogeneity in a highly homologous population of CF patients.

Resequencing studies have been successful identifying rare variants of possible effect within GWAS loci. A recent resequencing study for ulcerative colitis found many rare coding variants associated with the disease. The authors speculated that these findings might provide insight into the structure and function of the genes involved, perhaps guiding development of therapeutics (*65*). Unlike most re-sequencing studies, we opted to target the disease causing locus, rather than modifiers identified via CF GWAS. However, we and others are also pursuing these modifier loci in independent research projects.

It may not be intuitive to sequence *CFTR* in patients whom have already had genotyping sufficient to determine their causative alleles (e.g. F508del, G551D, R117, etc). Individuals in the U.S. are often screened at birth for cystic fibrosis through a blood spot test (IRT, see above) (*10*). If a newborn screens positive, additional testing will be conducted, including genotyping of the *CFTR* locus. This primary genotyping is usually conducted via a highly targeted array approach. In cystic fibrosis, this panel may contain anywhere from 23 to over 100 known CF mutations (out of >2000 variants in the gene). Most individuals with CF will have mutations in one or two of these most frequently occurring mutations. If two known disease causing mutations are detected, no further genotyping is usually required (*66*). Therefore, much of the variation within and around this locus (introns, untranslated regions, and intergenic regions) goes completely untyped in these patients. Newer diagnostic assays relying on "full" sequencing of the locus often do not characterize variation far beyond 100bp into the intronic regions (*67*).

In **Chapter 3**, I describe the extensive variation discovered in these uncharacterized regions in a large population of F508del patients, and how this variation may modulate disease traits.

Atypical Cystic Fibrosis

Sometimes, even after extensive genotyping of *CFTR*, patients with CF like phenotypes are found to carry no disease causing mutations at this locus. These patients are said to have 'atypical' CF or CF-related disorders. This distinct patient cohort was first recognized by Ron Rubenstein and Garry Cutting in 1998, "In approximately 2% of patients, there is an atypical phenotype, which consists of chronic sinopulmonary disease, pancreatic sufficiency, and either borderline (40 to 60 mmol/L) or normal (<40 mmol/L) sweat chloride concentrations"(*68*). These findings were based on both personal observations and previous case reports of patients with mild CF phenotypes and one or no *CFTR* mutations (*69-71*).

In his 2002 manuscript, Joshua Groman assessed *CFTR* mutations, function, and linkage in sibling sets and pedigrees with an atypical CF phenotype (*72*). He found that patients entering the study with one known mutation in *CFTR* were more likely to have a second mutation upon further testing. Genetically, these individuals have classic CF, but were phenotypically indistinguishable from the remainder of the cohort, which had either 1 or 0 *CFTR* mutations. Linkage studies of the *CFTR* locus in two pedigrees with elevated sweat chloride levels revealed *CFTR* was not causative of their atypical phenotype. Furthermore, CFTR specific chloride transport in the nasal epithelium was completely normal. This study was the first indication that genes beyond *CFTR* could

cause a CF like phenotype (genocopies). Later studies of the same patient cohort by Groman helped further characterize the atypical CF phenotype (*73*).

In 2005, Molly Sheridan genotyped three subunits (*SCNN1A ,SCNN1B,* and *SCNN1G*) of the epithelial sodium channel (ENaC) in an atypical CF cohort (*74*). It had been previously found that this protein is regulated via direct interaction with CFTR (*75, 76*). Additionally, other diseases with phenotypes overlapping atypical CF had reported mutations within subunits of this protein in affected patients. These include pseudohypoaldosteronism type 1 (PHA1) and Liddle syndrome. PHA1 is an autosomal recessive disorder characterized by renal salt wasting, elevated sweat chloride levels, hyponatremia, hyperkalemia, and elevated aldosterone levels. PHA1 was linked to mutations in the alpha, beta, and gamma ENaC subunits in 1996 by Chang (α,β) and Strautnieks (ϒ) (*77, 78*). PHA1 causative variants result in loss of ENaC function and salt-wasting, and thus these patients are treated with salt supplementation (*79*). Liddle syndrome is associated only with mutations in the beta subunit of ENaC (*SCNN1B*) (*80*). Unlike PHA1, these mutations result in a gain of ENaC function, and patients are treated with a low salt diet. In Sheridan's paper, she found that additional unreported mutations within *SCNN1B* were causative of an atypical CF phenotype (*74*). However, only 2 of the 20 patients studied carried causative variants, suggesting that additional unidentified loci may be causative of this phenotype.

As part of my thesis work, I have sought to elucidate possible causative genes for atypical CF phenotypes through exome and whole genome sequencing analysis. Many of the families originally reported by Molly Sheridan and Joshua Groman were included in these studies. While we have discovered many possible candidates in pathways related to

CFTR, few candidates currently have compelling genetic or biological evidence to warrant functional studies. In **Chapter 2**, I will discuss the discovery and molecular evaluation of variants within carbonic anhydrase 12 (*CA12*) in a family presenting with elevated sweat chloride levels and hyponatremic dehydration due to salt wasting. This is one of several exome sequencing analyses I conducted which resulted in determination of the molecular etiology of an atypical CF phenotype.

Next Generation Sequencing Analysis

Both chapters of this thesis heavily relied on a technology which steadily grew in popularity during my graduate studies: next generation sequencing (NGS). This massively-parallelized approach to genomics has had a profound influence on our ability to discover and understand variation in large populations, across cell and tissue types, and among species.

Chapters 2 and 3 utilize a targeted approach to sequencing regions of interest (as opposed to "shotgun" sequencing, often used for whole genomes). Very generally in this approach, a genomic DNA sample is first sheared into small fragments (~300-800bp). Next, a small adapter sequence is ligated to each end (~50bp). This adapter sequence contains a barcode and can be used to identify a given sample preparation. During capture, the prepared gDNA is hybridized with bait sequences (previously designed based on known regions of interest). These baits are often biotinylated such that, after hybridization with the DNA, they can be pulled down used streptavidin coated beads. This process allows one to then wash away the undesired DNA sequences. Following library preparation, the polymerase chain reaction (PCR) is sometimes used to amplify

the library. Finally, the library can be run through a flow-cell on an NGS platform such as Illumina's HiSeq. Illumina utilizes "sequencing by synthesis". In this process, fluorophore-tagged nucleotides are washed over the library which has been immobilized to the surface of a flow cell via the adaptor sequences and amplified. If a nucleotide is incorporated into a growing sequence of DNA ("synthesis"), it will emit light which is captured by the photodetector. This photodetector tracks each growing strand of DNA, monitoring which bases have been incorporated. The photodetector then conveys this signal to a computer, which processes the signal and generates a sequencing "read". The read contains two pieces of information: the ACTG sequence of the read, and a confidence score associated with that call. These reads can then be analyzed by bioinformatics software downstream (e.g. alignment, variant calling).

The majority of my studies have been devoted to developing and optimizing the post-sequencing processing of NGS data. There are three key steps in the process: read mapping or alignment, variant calling, and variant annotation. The methods which can be used at these steps are *highly* variable. The techniques employed must be carefully constructed and optimized to the type of assay (whole genome, exome, targeted, RNA, etc.), and the downstream analysis (research vs. clinical, desired sensitivity/specificity, filtering or burden testing).

Briefly, alignment is the process by which reads are mapped back to a known reference sequence. This is the most popular approach in human samples because the reference is well annotated, and the process is relatively quick. Alternatively, a *de novo* alignment can be performed where reads are aligned into contigs without the use of a reference. This method is slower and often employed in shot-gun sequencing and certain

types of RNA studies). There are many alignment software currently available, all of which have their pros and cons. In my studies, I have used "gapped" aligners, which allow for insertions and deletions within the reads (*81, 82*). The initial alignment is often optimized through a variety of steps including marking of duplicate reads and local *de novo* alignment in regions of high variability (often around insertions or deletions) (*83*).

Both small and large variants can then be called from the alignment (and are often informed by the unmapped reads). Variant callers are perhaps even more abundant than aligners, and again, each has its own pros and cons. There is marked discordance between calls from competing software (*84*). In order to reduce the fraction of false positive calls in my dataset, I have incorporated up to 5 variant callers into my pipelines. A final call set is then generated from the majority call (i.e. requiring 3 out of 5 software to be concordant for a given call). However, it is difficult to assess the accuracy of this final call-set, which may have a high false negative rate, given the lack of a consensus truth set at the time of its development. Finally variants can be annotated with meta-data such as conservation and allele frequency. This can be done using popular software (*85*), or using custom scripts in combination with public repositories.

The downstream analysis of these datasets is really where the NGS field is currently exploring and growing. Much of my thesis work has been spent writing scripts to correctly conduct these analyses. As time progresses, more methods will become available, allowing for more rapid and diverse processing of this data. Additionally, as these advanced analyses become more popular, a larger community will be available, hopefully resulting in the development of creative new approaches.

Next generation sequencing allowed me to pursue many scientific hypotheses during my studies – two of which are highlighted in this thesis. In **Chapter 2**, exome sequencing of one pedigree and one unrelated proband was employed in order to elucidate disease causing variants associated with their rare phenotype. In **Chapter 3**, a custom designed targeted resequencing capture was used to study variation *in cis* with the F508del allele in 762 cystic fibrosis patients.

Chapter 2

Loss of carbonic anhydrase XII function in individuals with elevated sweat chloride concentration and pulmonary airway disease

Briana Vecchio-Pagán; Melissa Lee; Neeraj Sharma; Abdul Waheed; Xiaopeng Li; Karen S. Raraigh; Sarah Robbins; Sangwoo T. Han; Arianna L. Franca; Matthew J. Pellicore; Taylor A. Evans; Kristin M. Arcara; Hien Nguyen; Shan Luan; Deborah Belchis; Jozef Hertecant; Joseph Zabner; William S. Sly; Garry R. Cutting. *Human Molecular Genetics* 2016. doi: 10.1093/hmg/ddw065**.**

Abstract

Elevated sweat chloride levels, failure to thrive (FTT), and lung disease are characteristic features of cystic fibrosis (CF, OMIM #219700). Here we describe variants in *CA12* encoding carbonic anhydrase XII in two pedigrees exhibiting CF-like phenotypes. Exome sequencing of a white American adult diagnosed with CF due to elevated sweat chloride, recurrent hyponatremia, infantile FTT and lung disease identified deleterious variants in each *CA12* gene: c.908-1 G>A in a splice acceptor and a novel frameshift insertion c.859_860insACCT. In an unrelated consanguineous Omani family, two children with elevated sweat chloride, infantile FTT, and recurrent hyponatremia were homozygous for a novel missense variant (p.His121Gln). Deleterious *CFTR* variants were absent in both pedigrees. CA XII protein was localized apically in human bronchiolar epithelia and basolaterally in the reabsorptive duct of human sweat glands. Respiratory epithelial cell RNA from the adult proband revealed only aberrant *CA12* transcripts and in vitro analysis showed greatly reduced CA XII protein. Studies of ion transport across respiratory epithelial cells in vivo and in culture revealed intact *CFTR*-mediated chloride transport in the adult proband. CA XII protein bearing either p.His121Gln or a previously identified p.Glu143Lys missense variant localized to the basolateral membranes of polarized MDCK cells, but enzyme activity was severely diminished when assayed at physiologic concentrations of extracellular chloride. Our findings indicate that loss of CA XII function should be considered in individuals without *CFTR* mutations who exhibit CF-like features in the sweat gland and lung.

Introduction

Persistently elevated sweat chloride concentration caused by loss of function mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene is the diagnostic hallmark of cystic fibrosis (CF). Individuals with features of CF who do not carry any disease-causing *CFTR* alleles have been reported. These patients were phenotypically indistinguishable from CF patients carrying two known CF-causing mutations (*72*). Some individuals presenting a milder, atypical CF were found to carry variants that altered the function of subunits that form the epithelial sodium channel (ENaC) (*74*),(*73*).

CA12, the gene encoding carbonic anhydrase (CA) XII, has been implicated as a cause of elevated sweat chloride concentration, failure to thrive in infancy, and recurrent hyponatremia in two consanguineous Bedouin kindreds (*86*),(*87*). The same missense variant was identified in both pedigrees. However, the variant caused only a modest reduction (~30%) in enzymatic activity (*87*), which was unexpected as autosomal recessive disorders are generally associated with severe loss of function variants. The authors speculated that the minimal reduction in CA XII function produced a phenotype limited to the sweat gland (OMIM #143860) (*87*),(*88*).

In this study, we report the discovery and analysis of loss of function variants in *CA12* that associate with elevated sweat chloride concentrations in two unrelated pedigrees. An adult proband in one pedigree also displayed pulmonary features that overlap with CF; namely recurrent pulmonary exacerbations, Pseudomonas in sputum cultures, and mild but distinct bronchiectasis upon high resolution chest CT scanning.

These findings indicate that loss of CA XII activity is uncompensated in certain epithelia and that CA XII may play a key role in the function of the pulmonary airways as well as the sweat gland.

Results

Identification of *CA12* variants segregating in two unrelated pedigrees

The proband in pedigree A (II:1, **Figure 2.1A**) presented with failure to thrive at 2.5 months of age and sweat chloride concentrations ranging from 82 to 88 mEq/L. She was diagnosed with cystic fibrosis (CF) and prescribed pancreatic enzymes to improve growth. At 7 months of age she had an episode of hyponatremic dehydration requiring hospitalization (plasma sodium 120 mm/L upon admission). Spirometry from ages 7-9 years indicate three episodes of airway obstruction, with forced expiratory volumes (FEV1) and forced expiratory flows (FEF25-75%) falling below 80%. At age nine, repeat sweat chloride testing revealed elevated levels (range=112-116 mEq/L) and serum IRT levels were within normal range, resulting in the discontinuation of pancreatic enzymes. Nasal potential difference (NPD) testing performed at this time reported aberrant chloride transport consistent with CF. It should be noted that NPD testing can have considerable technical variability and was standardized after this test was administered to proband A. Available clinical records between the ages of 19-22 revealed a persistent cough and cultures of bacteria common to CF patients, including Pseudomonas aeruginosa in the throat (age 19), Stenotrophamonas maltophilia in the throat (age 20), and Pseudomonas fluorescens in sputum (age 22). The proband's pulmonary exacerbations were often treated with a regiment consistent with her CF diagnosis, including bronchodilators,

antibiotics, and steroids. At age 25, NPD testing repeated at the same clinical facility was not consistent with CF (response to low chloride and isoproterenol: -24 mV on right and -16 mV on left). The proband continued to be seen regularly at an accredited CF care center and reported compliance with daily respiratory treatments including aerosolized albuterol, acetylcysteine, hypertonic saline, and chest physiotherapy. High resolution chest CT scanning revealed mild bronchiectasis without scarring, inflammation, or mucus plugging (**Figure 2.2**). Assessment of airway dilatation was confirmed by two adult CF pulmonologists and an additional interpretation by a radiologist who was masked to the clinical status of proband A. The proband has also been treated by a dermatologist for axillary hyperhidrosis. Exome sequencing was performed on the proband, her unaffected sister, and both parents. Average depth of coverage was 87X, and 93.2% of the targeted regions were covered at a depth ≥ 10X. No deleterious variants were found in *CFTR* or the three genes (*SCNN1A, SCNN1B,* and *SCNN1G*) that encode the epithelial sodium channel (ENaC). Loss of ENaC function can cause pseudohypoaldosteronism, a secondary and rare cause of elevated sweat chloride concentration (*74*),(*89*). Two variants within *CA12* were discovered in trans (compound heterozygosity) in the proband (II:1): a variant inherited from her father (I:1) in the canonical splice acceptor site of exon 10, c.908-1 G>A (rs148438059, chr15:63,619,433, ClinVar accession# SVC000255965) and an insertion variant of four nucleotides inherited from her mother (I:2), c.859_860insACCT (chr15:63,631,029-63,631,030, ClinVar accession# SVC000255963) in exon 8. The splice acceptor variant is found in heterozygosity in 53 individuals in the Exome Aggregation Consortium (ExAC) variant browser (*90*) with a global MAF of 0.00471%. It is the most common predicted deleterious *CA12* variant found in ExAC.

The proband's insertion mutation was not found in ExAC. *CA12* variants were confirmed in all family members via Sanger sequencing.

In a second unrelated family, a six year old Omani boy (proband B, II:3, **Figure 2.1B**) presented with a history of hyponatremic dehydration, elevated sweat chloride, and bilateral hyperkeratosis of the heels. Hyponatremic dehydration was alleviated with administration of Pedialyte and unrestricted access to dietary salt. Four sweat chloride measurements ranged from 90 to 110 mEq/L. Pulmonary function tests and fecal elastase measurements were within the normal ranges, ruling out chronic pulmonary and exocrine pancreatic insufficiency associated with CF. Aldosterone measurements excluded pseudohypoaldosteronism. Clinical diagnostic sequencing of the coding and intron flanking regions of *CFTR* and *SCNN1A* encoding the α subunit of ENaC in proband B did not detect sequence variations predicted to be deleterious. An 11 year old sister (II:1) of proband B initially considered to be asymptomatic was discovered to have a sweat chloride of 130 mEq/L. At a two year follow up, this sister was found to have developed a phenotype concordant with that of proband B, reporting episodes of hyponatremic dehydration as well as mild bilateral hyperkeratosis of the heels. Hyperkeratosis of the heels was not observed in proband A or the previously reported patients (*88*) and could be due to unrelated deleterious recessive alleles that may be present in this consanguineous pedigree. Unaffected siblings in pedigree B had sweat chloride measurements within the normal range for the referring laboratory (<50 mEq/L, personal communication Jozef Hertecant). No evidence of respiratory disease was reported in either patient; however, we were unable to obtain high resolution chest CT scans. Proband B was reported to have a normal chest X-ray at age 11 and his affected sister had

no pulmonary testing of any kind. *CFTR*, *SCNN1B*, and *SCNN1G* were excluded via genetic linkage analysis of all individuals in pedigree B with the assumption of recessive inheritance. Exome sequencing was conducted on proband B, his affected sister, and all unaffected siblings in pedigree B. The average depth of coverage was 34X, and 62.5% of the targeted regions were covered at a depth ≥ 8X. A previously unreported variant c.363 C>A (chr15:63,637,742, ClinVar accession# SVC000255964) in exon 4 of *CA12* was found in homozygosity only in the two affected individuals (**Figure 2.1B**). It is predicted to cause a substitution of His at codon 121 with Gln (p.His121Gln). Segregation of the *CA12* c.363 C>A variant in an autosomal recessive inheritance pattern was confirmed in all family members in pedigree B via Sanger sequencing.

CA XII is localized to the basolateral membrane of ductal epithelia in sweat gland and apical membrane in airway epithelia.

To ascertain whether CA XII was expressed in the organs affected in the two probands, namely the sweat gland and the airways, immunohistochemistry (IHC) of normal skin and lung sections was performed. IHC of whole skin tissue showed robust sweat gland expression of CA XII in the basolateral compartment of the reabsorptive ductal cells (**Figure 2.3B, 2.3C**). Basolateral CA XII staining in the reabsorptive sweat duct was distinguishable from apically localized CA II, a ubiquitously expressed cytosolic carbonic anhydrase (**Figure 2.3D, 2.3E**). To determine if CA XII is expressed in the airways, IHC of lung was performed and showed robust apical localization of CA XII in bronchiolar epithelia (**Figure 2.3F, 2.3G**). Of note, IHC of CA XII in the lung was performed using the same anti-CA XII antibody (ProteinTech #15180-1-AP) that was utilized for IHC of the sweat gland. The varying subcellular localization of CA XII

(basolateral within the sweat gland; apical within bronchioles), may indicate alternative roles for this protein in reabsorptive or secretory epithelial membranes.

The c.908-1 G>A and c.859_860insACCT variants found in proband A generate aberrant RNA transcripts

The two variants identified in proband A were predicted to affect RNA processing. To evaluate this supposition, respiratory epithelia from the inferior nasal turbinate was obtained for RNA and functional studies. CA XII is expressed in nasal epithelial cells (**Supplemental Table 2.1**) and respiratory epithelia of the nasal turbinates have been used as a proxy for respiratory epithelia of the airways (*91*). The *CA12* gene is composed of 11 exons that constitute its primary mRNA transcript (**Figure 2.4A**). Alternative splicing of *CA12* has been observed in native and cancerous tissues by both RT-PCR and RNA sequencing (*92*). The most common alternative isoform of *CA12* (CCDS# 10186) removes exon 9, a small exon composed of 33 bp, allowing the downstream transcript to retain the same reading frame. A review of publicly available RNA-sequencing splicing data from the Human Protein Atlas (*93*) reveals this isoform is predominately expressed in brain, and select other tissues (**Supplemental Table 2.1**). Two additional rare isoforms of *CA12* result from skipping of exons 9 and 10, or exon 10 only. These alternative *CA12* transcripts are of very low abundance in nasal and bronchial epithelial cells compared to the full-length transcript with 11 exons. PCR of cDNA derived from nasal epithelial cell RNA of proband A generated DNA products of 1069 bp, 980 bp, and 947 bp. Each product was gel purified and subject to Sanger sequencing. The 1069 bp product corresponded to full-length *CA12* transcript bearing the insertion c.859_860insACCT. This variant introduces a frameshift that is predicted to lead to the

incorporation of 49 novel residues following codon 287 and a premature termination codon (PTC) in exon 11 (predicted size 336 residues; **Figure 2.4B**). Despite the presence of a PTC, the transcript was stable due to the location of the PTC in the last exon of *CA12*, thereby allowing the transcript to evade nonsense mediated RNA decay (*94*). The splice site variant found in proband A, c.908-1 G>A, was predicted to cause misplicing of *CA12* exon 10 as it alters an invariant nucleotide of the canonical 3' splice acceptor site. Indeed, the 980 bp amplicon was *CA12* transcript missing exon 10 and the 947 bp product was an alternatively spliced *CA12* transcript missing exons 9 and 10 (**Figure 2.4C**). Loss of exon 10 was predicted to result in a frameshift beginning at codon 302 and translational read-through of the native termination codon. A novel termination codon in the 3' UTR occurs at amino acid position 413. The resulting protein is predicted to be composed of the first 302 amino acids of CA XII followed by 111 novel residues, 89 of which are translated from the 3' UTR. Skipping of exons 9 and 10 would add the same 111 novel residues but the frameshift would start at codon 291 (predicted size 402 residues). Finally, amplification from exon 8 to exon 10 and Sanger sequencing verified that all transcripts bearing exons 9 and 10 contained the c.859_860insACCT insertion (data not shown). In summary, all *CA12* mRNA transcripts in the nasal epithelial RNA of proband A were abnormal and each was predicted to generate aberrant CA XII protein.

<u>*CA12* variants found in proband A generate unstable CA XII protein</u>

Expression vectors with *CA12* cDNA modified to correspond to each of the three transcripts observed in the nasal epithelial cells of proband A, the missense variant p.His121Gln (c.363 C>A) observed in proband B, and a previously described Bedouin missense variant p.Glu143Lys (c.427 G>A) were transfected into HEK293 cells and

lysates were subjected to analysis by Western blot. Wild-type (WT) CA XII was present in both unglycosylated (39 kDa) and fully glycosylated (43 kDa) forms (*95*) (**Figure 2.5, lane 2**). CA XII protein was severely reduced in the lysate of cells transfected with *CA12* expression vectors bearing the insertion variant c.859_860insACCT (7.2%-13.3% of WT) and only a single band of the predicted mass of the unglycosylated protein (37.9 kDa) was observed (**Figure 2.5**, lane 4). CA XII lacking residues encoded by exon 10 and exons 9 and 10 were barely visible (**Figure 2.5**, lanes 5, and 6). CA XII bearing the missense variants p.His121Gln, and p.Glu143Lys generated protein of a molecular mass comparable to WT and unglycosylated and glycosylated forms were observed (**Figure 2.5**, lanes 7 and 8). CA XII with p.Glu143Lys had a higher fraction of unglycosylated protein, suggesting a possible effect of the amino acid substitution on processing and post-translation modifications. These results indicate that each of the changes in amino acid composition due to the *CA12* variants found in proband A cause substantial instability in CA XII.

Nasal respiratory epithelial cells from proband A demonstrate *CFTR*-mediated chloride transport

Cultured epithelial cells from proband A and controls were mounted in Ussing chambers for short circuit current measurements. To increase the driving force for chloride secretion through *CFTR*, the apical membrane was hyperpolarized by administration of amiloride that inhibits sodium current conducted by epithelial sodium channels. To specifically examine chloride currents mediated by *CFTR*, calcium-activated chloride channels were inhibited by DIDS (4, 4'-diisothiocyanoto-stilbene -2, 2'-disulfonic acid). Application of DIDS did not result in a significant change in current

in any sample. *CFTR* was activated by elevating cellular levels of cAMP with forskolin and 3-isobutyl-1-methylxanthine (IBMX). Upon treatment with forskolin and IBMX ("F+I"), the change in *CFTR*-mediated chloride transport in nasal epithelia from proband A (9.34 and 9.15 uA/cm2) were higher than that observed in nasal epithelial from a CF subject tested concurrently (2.1 uA/cm2). The values in proband A are consistent with short circuit measures of nasal epithelia from other non-CF and CF subjects (*96*). Substantial reduction in the current of cells from proband A upon addition of GlyH-101 (-18.8 and -21.5 uA/cm2) is consistent with the chloride secretion being mediated by *CFTR*. Together, these findings suggest that loss of CA XII does not ablate *CFTR*-mediated chloride secretion across nasal respiratory epithelia.

<u>p.His121Gln and p.Glu143Lys mutations cause near complete loss of enzyme activity of CAXII.</u>

As the missense variants permitted the generation of stable full-length protein, immunocytochemistry and confocal microscopy was utilized to test whether either variant affected the subcellular localization of CA XII. Expression in polarized epithelial Madin-Darby canine kidney (MDCK) cells revealed that WT CA XII localized to basolateral membranes (green, n=7, **Figure 2.6**, left panel). CA XII bearing p.His121Gln (green, n=8, **Figure 2.6**, middle panel) and p.Glu143Lys (green, n=13, **Figure 2.6**, right panel) showed basolateral localization indistinguishable from that of WT CA XII. Comparable staining patterns were observed when immunocytochemistry was performed with alternative CA XII antibodies (rabbit: Sigma Prestige; mouse: Novus) that detected different extracellular CA XII epitopes (data not shown). Since the p.His121Gln and p.Glu143Lys mutants were localized to plasma membranes, we tested the effect of each

variant on CA function by measuring carbonic anhydrase enzyme activity. Carbonic anhydrase activity is the reversible rate of CO2 hydration. When assayed in a 2 mM NaCl solution, the activity of CA XII bearing p.His121Gln is reduced by 84.6 ± 3.6% compared to WT (n=11) while the enzymatic activity of CA XII with p.Glu143Lys is reduced by 24.4 ± 4.9%, consistent with the approximate 30% reduction in activity previously reported under the same conditions (*87*) (**Figure 2.7**). When assayed at physiological salt concentrations, the enzyme activity of p.His121Gln was reduced by 99.2 ± 0.5% compared to WT, and the activity of the p.Glu143Lys mutant was reduced by 97.1 ± 1.2% compared to WT (n=9) (**Figure 2.7**). The enzyme activities of p.His121Gln and p.Glu143Lys were not statistically different (p = 0.12; t test) when assayed in the presence of 100 mM NaCl. These findings reveal a chloride-sensitive abolition of carbonic anhydrase enzyme activity for CA XII p.His121Gln and p.Glu143Lys mutants compared to WT.

Discussion

Each of the *CA12* variants reported in the three individuals described here with elevated sweat chloride, recurrent hyponatremia and failure to thrive in infancy cause severe loss of CA XII activity. The two variants found in proband A generate mRNA transcripts that are missing nucleotides that form transmembrane domains and enable dimerization via a key glycine zipper motif (*95*). Translation of cDNA that replicate the mRNA transcripts identified in proband A generated unstable protein products in heterologous cells. On the other hand, CA XII proteins bearing the missense variant found in this study and the previously reported missense variant were stable and, in each case, localized to the basolateral membrane of MDCK cells, consistent with native CA

XII location in kidney cells (*97*). Point mutation energy modeling by FoldX suggested that both missense variants should affect CA XII structure and catalytic function. Modeling indicated that the secondary amine of histidine 121 is essential for tetrahedral coordination of the zinc ion within the catalytic domain. In the WT conformation, the H121–zinc bond distance is ~2.1Å (**Supplemental Figure 2.1**). When mutated to glutamine as in p.His121Gln, the distance from zinc to the hydroxyl of the carboxyl group is 3.042Å. This increased distance likely precludes formation of a coordinating bond, and the zinc ion is instead coordinated by a free hydroxide ion (red sphere). In the extracellular aqueous environment, this coordination would only be transient, potentially leading to poor catalytic activity. The distances of other residues, such as the highly conserved second shell glutamic acid 143, are also altered by p.His121Gln by a magnitude similar to that reported in previous studies of the Bedouin p.Glu143Lys mutation (*86*). CA XII bearing the p.Glu143Lys mutation was particularly sensitive to inhibition by chloride, as previously reported (*87*). Given that the active site of CA XII lies on the extracellular face of the basolateral compartment in the sweat duct, enzymatic activity of both mutants was assayed in the presence of increasing NaCl concentrations. The concentration of NaCl which most closely mimics the enzyme's native physiological environment is 100 mM NaCl (*98*). At this concentration, catalytic activity of CA XII bearing each missense mutation was reduced to less than 3% of WT activity. As the affected individuals are homozygous for the *CA12* missense mutations, it is reasonable to conclude that the sweat gland dysfunction observed in each is due to near complete loss of CA XII activity. This conjecture is supported by the studies of proband A, where mutations in each *CA12* gene lead to severe instability of CA XII protein.

Robust expression of CA XII in lung epithelia and the observation of bronchiectasis in proband A suggest a role for this protein in the maintenance of airways. Elevated sweat chloride concentration indicates aberrant chloride transport in the sweat duct and is a consistent feature of the three individuals carrying the loss of function *CA12* variants reported here, as well as in 11 individuals homozygous for the p.Glu143Lys mutation reported previously (*88*). However, *CFTR*-mediated chloride transport appeared to be intact in the nasal respiratory epithelia as determined by in vitro and in vivo methods, suggesting that other pathways of ion transport in the airways might be disrupted by the loss of CA XII. Given the importance of CAs in the maintenance of pH via bicarbonate metabolism, the mechanism underlying airway damage could be related to aberrantly low pH of airway surface liquid (ASL) due to loss of bicarbonate production or transport. The pH of the ASL has been shown to be integral to the proper expansion and processing of mucins which play a key role in CF-related bronchiectasis (*99*). Alternatively, the bronchiectasis observed in proband A might be unrelated to the loss of CA XII function. Spirometry measurements in all tested individuals homozygous for the p.Glu143Lys mutation were reported as normal (*88*) and chest X-rays of proband B were reported to be normal at age 11; however, lung function measures and chest X-rays were also normal throughout the life of proband A up to her current age of 25. Detection of abnormal airway dilatation in proband A required high resolution CT scanning; therefore, it is possible that bronchiectasis remains undetected in the previously reported patients with loss of CA XII function and the affected individuals in pedigree B. Without comparison chest CTs from the affected individuals in pedigree B, it cannot be determined if proband B and his affected sister manifest bronchiectasis similar to that

observed in proband A. Further, if loss of CA XII function does lead to bronchiectasis, the possibly ameliorating impact of CF-specific airway treatments is an important question. Proband A has undergone routine airway clearance, therapies, and antibiotic courses since being diagnosed with CF as an infant. It is possible that a lifetime of diligent pulmonary monitoring and treatments managed by an accredited CF care center minimized the effect of bronchiectasis upon pulmonary function. However, estimating the impact of respiratory therapy on degree of airway dilatation is difficult without study of additional CA XII-deficient individuals manifesting pulmonary disease. Indeed, the identification of additional individuals with loss of CA XII function and high resolution chest CT scanning will clarify the role of CA XII in the airways. Given the strong causative connection between aberrant chloride transport and bronchiectasis in CF, CA XII loss of function should be considered as a potential explanation for non-CF bronchiectasis.

Loss of CA XII function in a patient with respiratory disease in humans suggests a previously unsuspected role for this isozyme in the lung. Although each individual CA isozyme follows a tissue-specific expression pattern (*100, 101*), RNA expression studies show that multiple isozymes can be expressed in certain tissues (*102*). Loss of function of one isozyme may be compensated for by other isozymes in certain tissues. This explanation is offered for the lack of a phenotype in patients with loss of function mutations in CA I (*103*). Our findings show that isozyme redundancy does not compensate for loss of CA XII function in the sweat gland. Further, although other transmembrane CA isozymes such as CA IV have been localized to the plasma face of

lung microcapillaries, our results suggest that CA XII may also play an important role in the maintenance of the airways.

IHC of the sweat gland revealed CA XII to be highly expressed in the resorptive duct in basolateral distribution consistent with its location in other epithelia, namely endometrium, kidney, and large intestine (*104*), (*105*), (*97*). Faint staining was observed at the apical membrane which may be non-specific signal or evidence of dual CA XII localization. However, the pattern of CA XII staining was distinctly different for that of CA II, which was discretely localized near apical membranes of the ductal cells. CA XII was only found on basolateral membranes of polarized MDCK cells. In contrast, CA XII was discretely localized to the apical regions of normal airway. The distribution of CA XII does not appear to be due to non-specific signal as two antibodies that detect different antigenic regions of CA XII revealed the same immunolocalization pattern. Localization to the terminal bar is consistent with CA XII location in other tissues, including the bronchus and fallopian tubes (*93*). A possible factor in the different localization of CA XII could be its involvement in one of the transport metabolon complexes formed between CA isozymes and bicarbonate transporters that facilitate the exchange of bicarbonate across membranes (*106*). Bicarbonate transport metabolons comprised of CA isozymes and anion exchangers have so far been described for three of the transmembrane isozymes: CA IV associates with AE1 (*106, 107*), CA IX associates with AE2 (*108*), and CA XIV associates with AE3 (*109*). AE1 and CA IV have been localized to both basolateral and apical membranes in different cell types in the kidney (*107*). Since the kidney can absorb and secrete ions including bicarbonate, it is possible that the different localization of the AE1/CA IV metabolon may be related to the direction of ion

flow. However, to date, no metabolon interaction has been reported for the remaining transmembrane isozyme, CA XII.

Our study suggests a mechanism for the well-established salt wasting complication of topiramate, a CA inhibitor and anticonvulsant commonly prescribed to people suffering from epilepsy. Elevated sweat chloride concentration is an established phenomenon observed in epileptic children being treated with topiramate (*110*). In these children, CF was clinically and/or molecularly excluded as being the cause of this increase in sweat chloride value, and the effect disappeared when topiramate treatment ended. Investigations into the inhibition potency of topiramate across the α family of CAs have shown that topiramate is a strong inhibitor of CA XII, and not the other transmembrane isozymes CA IV, IX, and XIV (*111*),(*112*). These pharmacologic observations are consistent with our hypothesis regarding the role of CA XII in the maintenance of proper ion composition in the sweat gland. Topiramate is also a strong inhibitor of the ubiquitous and highly active cytoplasmic isozyme CA II. Correspondingly, individuals being treated with topiramate have been observed to develop renal tubular acidosis (*113*), a hallmark feature of CA II deficiency syndrome.

In summary, the individuals studied here demonstrate that severe loss of function mutations in *CA12* cause an autosomal recessive disorder affecting chloride and sodium resorption in the sweat duct. The observation of airway dilatation in proband A suggests a possible molecular etiology for some forms of non-CF bronchiectasis, a disease that affects over 110,000 individuals in the U.S. (*114*).

Materials and Methods

Recruitment

Family A was recruited and consented into the study Genome-wide Sequencing to Identify the Genes Responsible for Mendelian Disorders at Johns Hopkins University (IRB# NA_00045758). Family B was referred to Johns Hopkins University via private communication of Dr. Jozef Hertecant at Tawam Hospital, United Arab Emirates. All members of family A were consented into the Molecular Genetics of Cystic Fibrosis (IRB# NA_00050260).

Linkage exclusion assays

Three highly polymorphic deCODE STRs were selected for each locus to be excluded (*CA12*, and *SCNN1B* and *SCNN1G*, which lie in close proximity to one another) from the STS Marker track on the UCSC Genome Browser. Each marker was no more than 1Mb from either end of the locus to be excluded. Primer sequences from the STS Marker track were run through BLAT to verify specificity. Oligonucleotides were synthesized by IDT. Forward primers were fluorescently labeled with 6-FAM. STRs were PCR amplified from genomic DNA and amplification products were separated on an ABI Prism 3100 Genetic Analyzer by automated capillary electrophoresis. Fragment sizing and visualization were performed using ABI GeneMapper software. Haplotype phasing was performed by manual inspection. If the proband and at least one unaffected sibling were found to be IBD2 for a locus, it was deemed to be excluded, given an assumption that the disease phenotype follows an autosomal recessive Mendelian inheritance pattern.

DNA sample acquisition, exome sequencing, and *CA12* genotyping

Peripheral blood was obtained from all consented individuals in pedigree A and genomic DNA was extracted by a phenol/chloroform protocol. Exome capture was performed on all siblings within the pedigree using the Agilent SureSelect Human All Exon (51 Mb), and 100 bp paired-end reads were subsequently obtained from an Illumina HiSeq 2500 system as part of a study within the Baylor-Hopkins Center for Mendelian Genomics conducted by the Center for Inherited Disease Research. Reads were aligned to the hg19 reference genome using Burrows-Wheeler Aligner software (*115*) and subsequent alignment processing was completed using SAMtools (*116*), PicardTools, and GATK softwares (*117, 118*) in a manner similar to (*119*). Filtering of variants was conducted via custom scripts, and variant prioritization was conducted using Enlis Genomic Research software. To verify mutations, a 654 bp region encompassing c.908-1 G>A and c.859_860insACCT was amplified from genomic DNA from the proband A and her mother by PCR using the following primers (IDT): 5'F GCCCTGTACTGCACACACAT and 3'R AGGATGATGCCCAGACTCAG. PCR products were purified using QIAquick PCR purification kit (Qiagen), and then sequenced using the Applied Biosystems 3730xl DNA Analyzer. The resulting sequences were analyzed via the Sequencher analysis suite (Gene Codes). Genomic DNA extracted from peripheral blood was obtained from all consented individuals in pedigree B. Exome capture was performed on all siblings using the Illumina Truseq Exome Enrichment kit (62 Mb), and 90 bp paired-end reads were subsequently obtained from an Illumina HiSeq 2000 system (Otogenetics). Exome sequencing data analysis was conducted in a manner similar to pedigree A, however, variant prioritization was conducted using VAAST

software (*120*). *CA12* (RefSeq# NM_001218.4) mutations and mode of inheritance were verified via Sanger sequencing. A 475 bp region encompassing c.363 C>A was amplified from genomic DNA of all siblings in pedigree B by PCRs using the following primers (IDT): 5'F GTCCCATGCTCTGGTGTTATC and 5'R CTTTCCAAGGTGAACCAAGAA. PCR products were purified, sequenced, and analyzed as described for pedigree A. The resulting sequences were analyzed via the Sequencher analysis suite (Gene Codes).

It should be noted that all variant nomenclature is specific to *CA12* nucleotide and CA XII amino acid numbering and represent the minus strand sequence unless otherwise specified. All genomic coordinates are specific to hg19. All variants discovered in this study have been submitted to ClinVar.

IHC of CA XII in human sweat duct and lung

Frozen discarded unidentified skin and lung obtained from the Division of Surgical Pathology, Johns Hopkins Hospital, Baltimore, MD, were embedded in Optimal Cutting Temperature (OCT) compound and held at -70°C prior to sectioning. Six μm cryo-sections were mounted onto uncoated microscope slides. Staining with H&E (Sigma, St Louis, MO, USA) for 1 min was performed for morphological evaluation. The rest of the slides were stored at -70°C until use. Sections were fixed for 10 min in pre-cooled acetone followed by 5 min peroxidase block at room temperature to quench the endogenous peroxidase activity. Sections were further incubated in serum-free protein block (Dako # X0909) for 20 min at room temperature and incubated overnight at 4oC with the following primary antibodies: anti-rabbit *CA12* (ProteinTech # 15180-1-AP) and

anti-rabbit CA2 (LSBio # C138796). Relevant universal negative control antibodies: mouse (Dako # N1698) and rabbit (Dako # IR600) were used to ascertain nonspecific staining. After washing, staining was performed using EnVision+System-HRP (AEC) kit from Dako (# K4008). Sections were covered with peroxidase-labeled polymer for 30 mins. For visualization of the reaction, sections were developed in AEC+substrate-chromogen for 5-20 mins. After washing, the sections were counterstained with hematoxylin (Dako # S3309) for 30 seconds, cleared, and mounted on Faramount aqueous mounting medium (Dako # S3025). Samples were analyzed under an Olympus BX51 microscope.

Nasal epithelial culture and Ussing chamber studies

Nasal epithelia from the proband of pedigree A were expanded and cultured using previous described method (*121*). Cells were mounted in Ussing chambers and studied as previously described (*122*). Apical and basolateral chambers contained the same bathing solution with symmetrical Cl- concentrations. *CFTR*-mediated Cl- current were measured using a previously described protocol (*122*).

Identification of *CA12* mutant transcripts

RNA was isolated from expanded nasal brushings of proband A by standard Trizol-chloroform method. cDNA was made using RT-PCR (Qiagen iScript) and served as template for *CA12* amplification. PCR products of *CA12* transcript isoforms were separated by 3% agarose gel electrophoresis and subject to Sanger sequencing.

Development of mutant *CA12* expression vectors

Full length wild-type (WT) *CA12* cDNA in bacterial pBS II expression vector was generously provided by Dr. William Sly. The proband B variant c.363 C>A (p.His121Gln) was introduced into full length WT *CA12* cDNA in bacterial pBSII expression vector using the QuikChange II XL Site-Directed Mutagenesis kit reagents and protocol (Agilent). Mutagenesis products were confirmed by Sanger sequencing. Each cDNA was removed from the bacterial expression vector using restriction enzymes KpnI and BamHI, purified by gel electrophoresis, and ligated into the eukaryotic pcDNA5 FRT expression vector. Subcloning was confirmed by Sanger sequencing. Proband A variant c.859_860insACCT was introduced into pcDNA5 FRTexpression vector bearing *CA12* cDNA using the QuikChange II XL Site-Directed Mutagenesis kit reagents and protocol (Agilent). A previously described *CA12* missense variant, c.427 G>A (p.Glu143Lys), deemed a moderate hypomorph and reported in a large consanguineous Bedouin kindred manifesting a highly similar phenotype (*87*), (*86*) was also introduced into the full length *CA12* cDNA. To replicate the consequences of the splice acceptor variant c.908-1 G>A found in proband A, *CA12* cDNAs lacking exon 10 and lacking exons 9 and 10 were custom synthesized (GeneWiz, South Plainfield, NJ). Since exon 10 skipping was predicted to result in a frameshift and subsequent readthrough of the natural termination codon, 287 bp of the *CA12* 3' UTR were included in each construct. Both constructs were removed from the bacterial pUC57 expression vector using restriction enzymes KpnI and EcoRV, purified by gel electrophoresis, and ligated into eukaryotic pcDNA5 FRT expression vector. Subcloning was confirmed by Sanger sequencing.

Expression of WT CA XII and mutant proteins in HEK293 cells

HEK293 cells were transiently transfected with 500 ng of WT and mutant *CA12* vectors using Lipofectamine 2000 Reagent and standard protocol (Invitrogen). Cells were lysed 24 hours post-transfection. Western blotting of cell lysates was performed using anti-CA XII antibody (Novus #NBP1-81668), and loading control GAPDH antibody (Sigma #G9545).

Immunocytochemistry of CA XII in a polarized epithelial cell line

Madin-Darby canine kidney (MDCK) cells were transiently transfected with 1.6 µg of *CA12* cDNA using 3.2 µl Lipofectamine 2000 Transfection Reagent and protocol (Invitrogen). Cells were fixed one day post-transfection with 4% paraformaldehyde for 20 mins and rinsed with 1X PBS. Cells were permeabilized with 0.5% Triton X-100 for 5 mins, then rinsed with 1X PBS, and blocked overnight at 4C with 2.5% goat serum. Cells were immunostained using rabbit anti-CA XII primary antibody (ProteinTech #15180-1-AP) diluted 1:200 and anti-ZO1 primary antibody (Invitrogen) with a conjugated anti-mouse red fluorophore diluted 1:200, followed by incubation in anti-rabbit secondary antibody diluted 1:50. Findings were validated by staining with different primary anti-CA XII antibodies from Sigma Prestige (mouse) and Novus (rabbit). All antibodies were diluted in 2.5% goat serum blocking solution. Cells were washed in 1X PBS three times for 10 mins following the 90 mins primary antibody incubation. Cells were washed in 1X PBS four times for 15 mins following the 30 mins secondary antibody incubation. Cells were mounted on microscope slides with Molecular Probes ProLong Gold Antifade

Reagent with DAPI and viewed with the Zeiss LSM510-Meta single-point confocal laser-scanning microscope and Zen imaging software.

Carbonic anhydrase enzyme activity assay

Cell pellets were lysed by sonication in 300 µl lysis buffer (PBS containing protease inhibitors, 1 mM each of PMSF, o-phenanthroline, EDTA, benzamidine-hydrochloride and iodoacetamide plus 1% NP-40) and left on ice. The media were centrifuged to remove dead cells. The protein concentration of cell lysates was determined by microLowry's procedure using bovine serum albumin as a standard (*123*). The carbonic anhydrase activity was determined using Maren's procedure (*124*) as described (*125*). To reflect extracellular physiological chloride concentration (*98*), 100 mM NaCl was utilized.

Acknowledgements

This study would not be possible without the participation of the patients and families described in this manuscript. The authors would like to acknowledge Diane Acquazzino for acquisition of patient records. This work was supported by the National Institutes of Health [R01 DK044003]; and the Cystic Fibrosis Foundation [CUTTIN13A2].

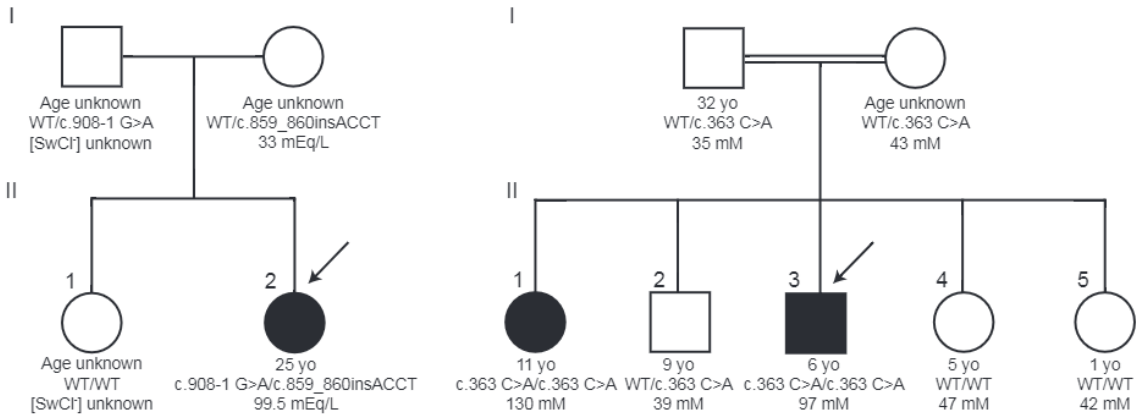


Figure 2.1. Segregation of putative deleterious CA12 variants in two unrelated families. Filled shapes indicate status as affected and arrows indicate the proband in each family. The ages indicate the age of the individual at the time of exome sequencing. (A) Pedigree A: A white American family in which the proband exhibits consistently elevated sweat chloride concentration and bronchiectasis. The proband carries a splice acceptor variant c.908-1 G>A and an insertion frameshift variant c.859_860insACCT. (B) Pedigree B: An Omani family with first-cousin parents as indicated by the double horizontal line. The proband and affected sister exhibit elevated sweat chloride concentrations and have experienced multiple episodes of hyponatremic dehydration. Only the proband and affected sister are homozygous for c.363 C>A (p.His121Gln).

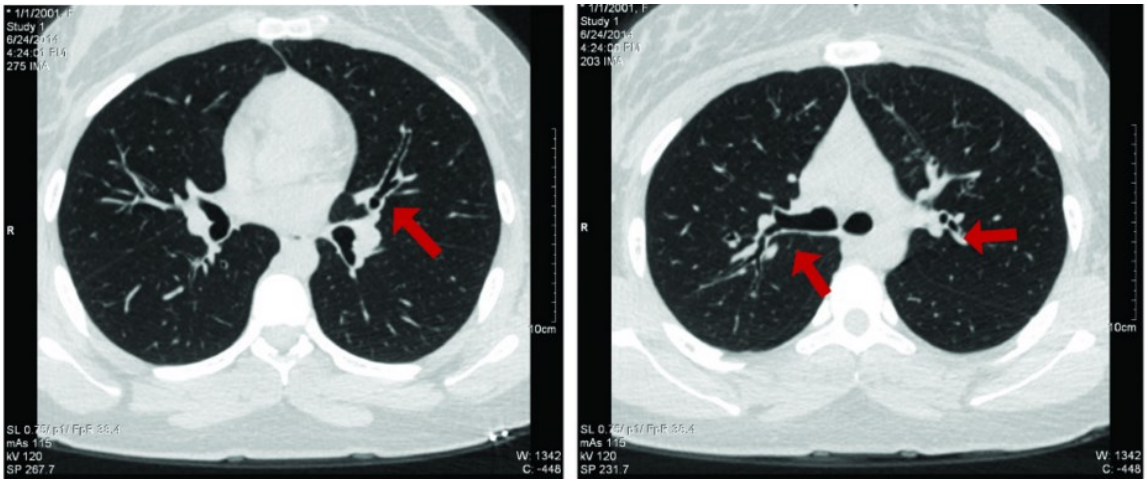


Figure 2.2. Axial plane high resolution CT images of proband A. Examples of enlargement of the airways (tram-tracking and signet rings) are indicated by red arrows. This bronchiectasis is seen in the absence of mucus plugging, scar tissue, or surrounding inflammation. It is unknown as to whether this mild pulmonary phenotype would be more exacerbated had the proband not been undergoing daily preventative lung therapies (aerosolized albuterol, hypertonic saline, chest physiotherapy, etc.) due to her original CF diagnosis.

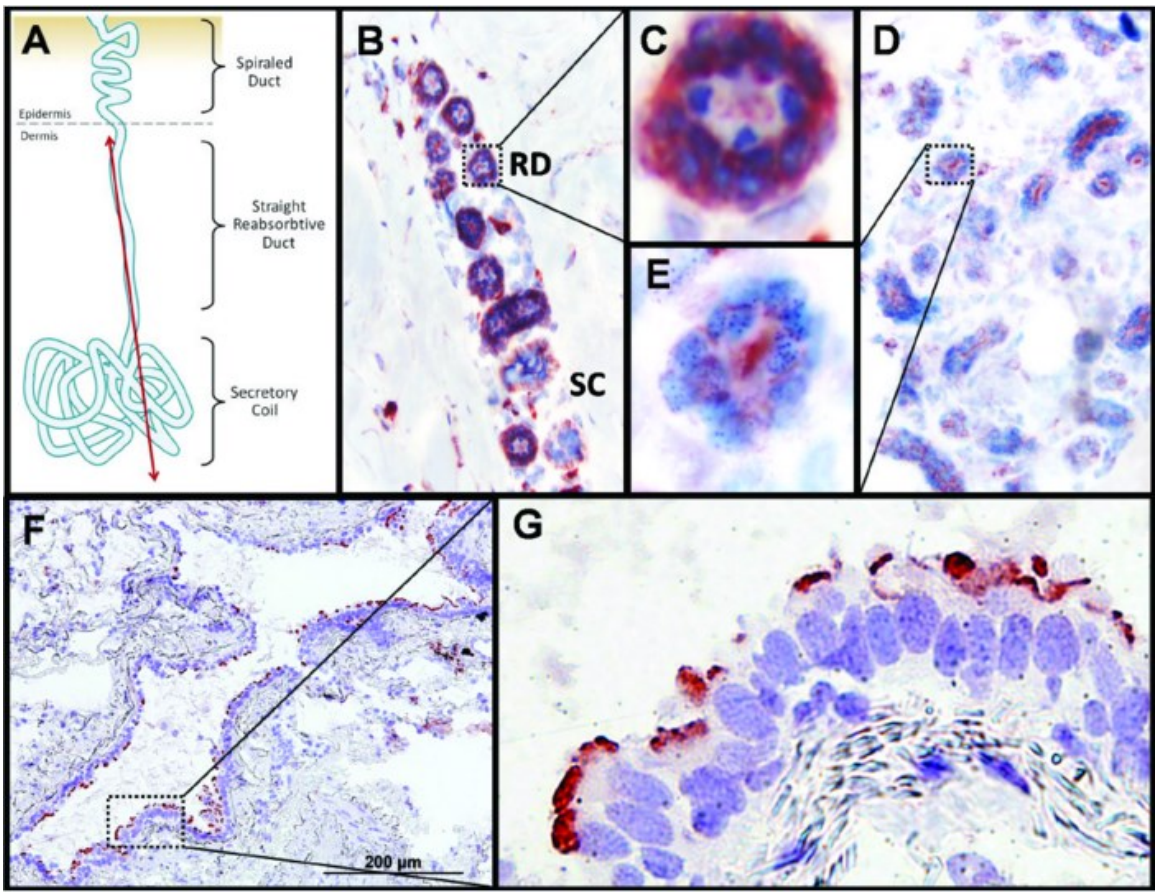


Figure 2.3 Immunohistochemical staining of CA XII and CA II in human skin and lung. (A) Diagram depicting longitudinal view of sweat gland components with red line indicating a hypothetical plane used to generate slices for the micrographs shown in panels B, C, D, E. (B) Two populations of sweat gland resorptive ducts ("RD") and secretory coils ("SC") immunostained for CA XII and counter-stained with hematoxylin and eosin. The resorptive ducts (n=9 different cross-sections captured in this panel) show positive staining for CA XII in a two cell thick layer of cuboidal epithelia cells. The myoepithelial cells surrounding the secretory coils (n=3 different cross-sections captured in this panel) show light staining of CA XII that may be non-specific. Magnification of this micrograph is 100x. (C) Enlargement of CA XII positive staining of the basolateral membrane in resorptive ductal cells from panel B. (D) Positive staining of apically

localized control protein CA II in sweat ducts and secretory coils. Magnification of this micrograph is 100x. (E) Enlargement of resorptive ducts from panel E showing apical localization of CA II. (F) CA XII positive staining of the luminal edge of a bronchiole cross-section. Magnification of this micrograph is 100x. (G) Positive staining of apically localized CA XII in the terminal bar of bronchial epithelia. Magnification of this micrograph is 400x.

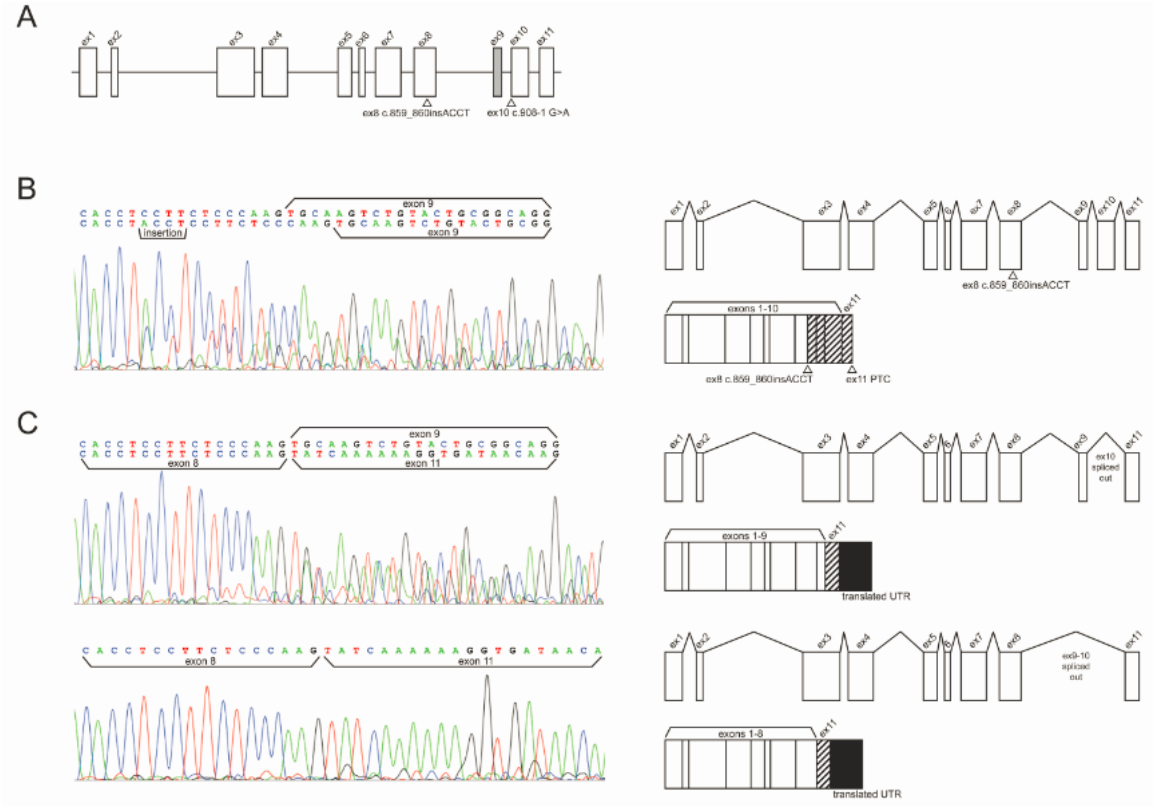


Figure 2.4. Effect of CA12 variants upon RNA processing in nasal epithelial cells from proband A. (A) Exon and intron structure of CA12 with locations of proband A variants identified by NGS. Rectangles represent exons and the lines through the center of the rectangles represent the genomic axis. Variants are indicated with triangles and HGVS cDNA names. The gray rectangle indicating exon 9 is spliced out in an alternative CA12 transcript whose function and tissue distribution is unknown. (B) (Left) Electropherogram of Sanger sequencing of cDNA reverse transcribed from RNA extracted from proband A cultured nasal epithelial. Sequencing detected a heterozygous insertion variant c.859_860insACCT on a transcript retaining alternatively spliced exon 9. A second transcript detected by sequencing does not bear the insertion, retains exon 9, and is consistent with transcript lacking exon 10 as a result of the in trans variant c.908-1 G>A. Presence of this second transcript in this sequencing reaction is likely due to

imperfect isolation of the similarly sized transcripts by gel purification as transcript bearing the insertion is only 89 bp longer than transcripts lacking exon 10 only. (Right) Gene models depict RNA processing of the insertion variant and the predicted gene product. The insertion variant causes a frameshift starting in exon 8, a premature termination codon in exon 11, and was predicted to generate a misfolded protein targeted by ERAD. Hashed rectangles indicate an altered exonic reading frame. (C) (Left) Electropherogram of Sanger sequencing detecting transcript missing exons 9 and 10, and a second transcript missing exon 10 only, due to heterozygous splice acceptor variant c.908-1 G>A. Imperfect isolation of transcripts in this reaction is due to alternatively spliced exon 9 which is only 33 bp long. (Right) Gene models depict the two processed RNAs and protein products lacking exon 10 observed by RT-PCR: one with exon 9 alternatively spliced out and one retaining exon 9. This variant is predicted to cause skipping of exon 10, a frameshift resulting in readthrough of the native stop codon, translation of 267 nucleotides from the 3' UTR, and a misfolded protein targeted by ERAD. Hashed rectangles indicate an altered exonic reading frame and filled rectangles indicate translation of the 3' UTR.

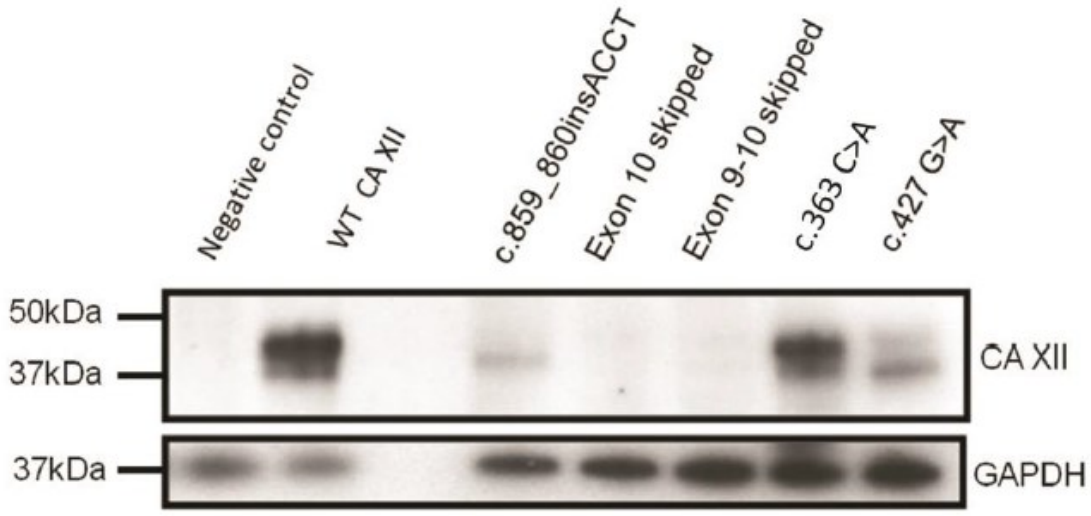


Figure 2.5. Expression of transiently transfected wild-type and mutant CA XII protein in HEK 293 cells. Western blot of cell lysates extracted from HEK 293 cells (top) following transfection with CA XII expression vectors. Probing with anti-CA XII antibody (Novus) shows unglycosylated (39 kDa) and glycosylated (43 kDa) protein generated from transfections with WT CA12 (lane 2), c.363 C>A (lane 7) and c.427 G>A (lane 8). The insertion variant in proband A (c.859_860insACCT) produced a faint band of approximately 38 kDa while CA XII cDNA missing exon 10 and exons 9 and 10 sequence generated only faint bands. The negative control in the first lane is a mock transfection. The third lane has no lysate. GAPDH loading control (bottom) shows loading of cell lysates.

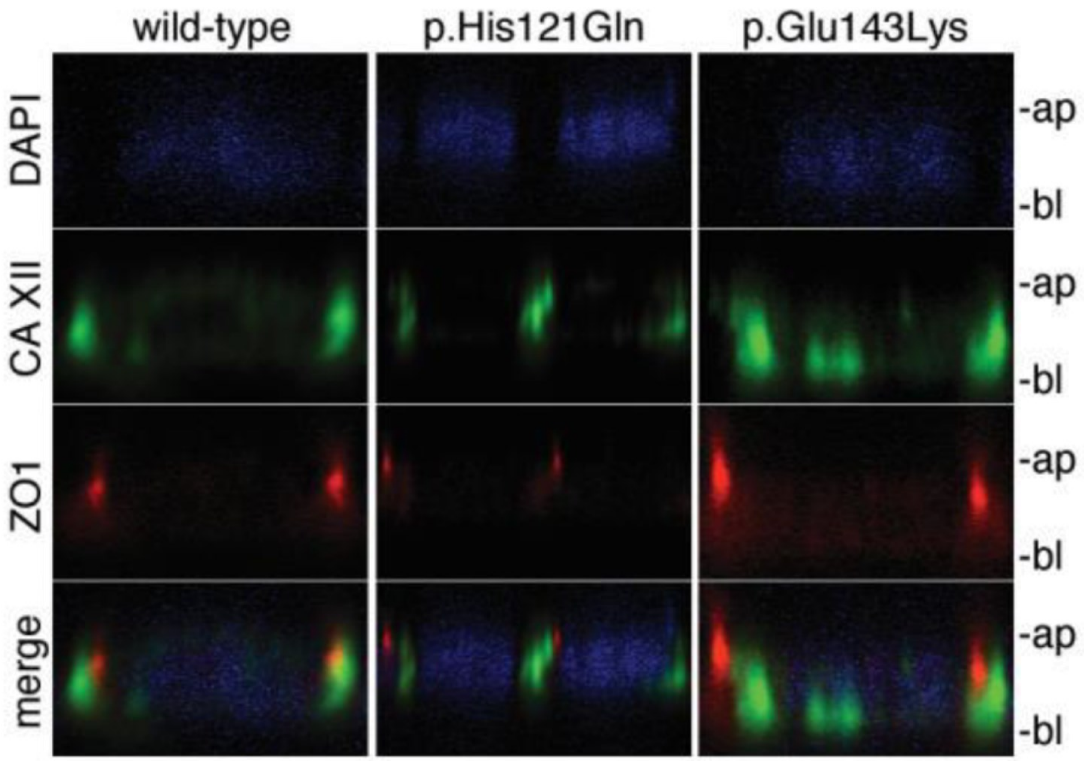


Figure 2.6. Subcellular localization of WT and mutant CA XII in polarized MDCK cells. Fluorescent co-staining of (left) WT CA XII (green), (center) p.His121Gln (green), and (right) p.Glu143Lys (green) with endogenous tight junction protein ZO1 (red) and nuclear stain DAPI (blue) in polarized MDCK cells imaged in the xz-plane. This micrograph reveals primarily lateral staining of CA XII; however, basal and lateral staining were observed for WT (n=7 different micrographs), p.His121Gln (n=8 different micrographs), and p.Glu143Lys (n=13 different micrographs). The apical membrane is indicated by "ap" and the basal membrane is indicated by "bl."

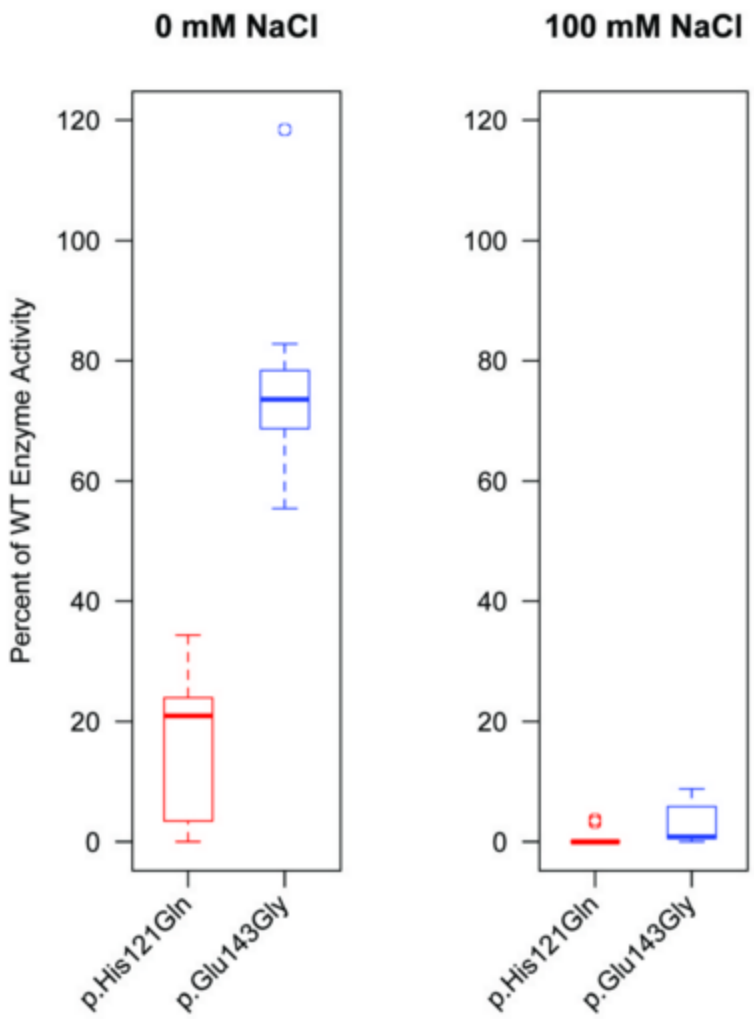
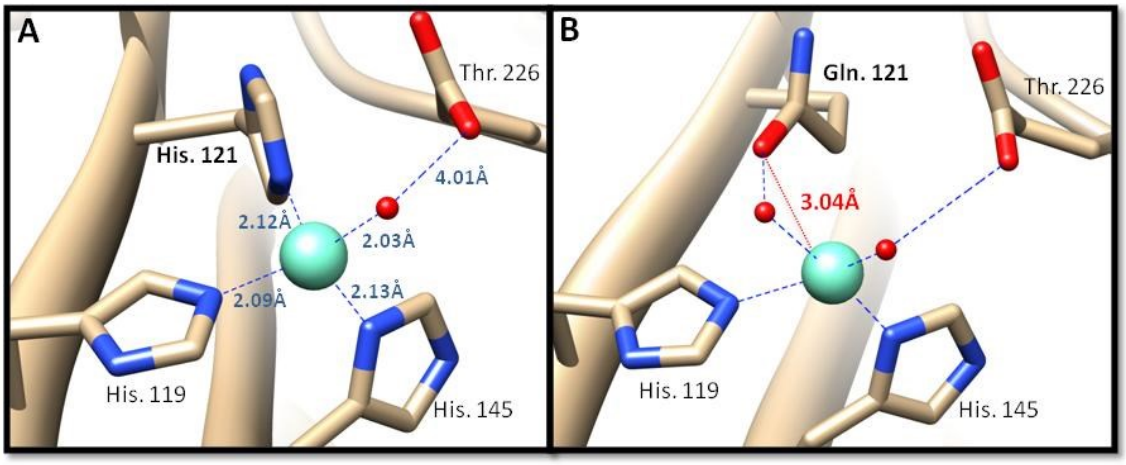


Figure 2.7. Enzymatic activity of CA XII proteins bearing p.His121Gln or p.Glu143Gly substitutions. Enzymatic activity of CA XII mutants p.His121Gln (red boxplot series) and p.Glu143Gly (blue boxplot series) was determined by assaying the reversible rate of hydration of CO2 as previously described in the absence (left) or presence (right) of physiological concentration of NaCl and normalizing to wild-type. Boxes represent the interquartile range (IQR) and the horizontal bars within the boxplots represent the median. The top whisker represents the 75th percentile plus 1.5 times the IQR and the bottom whisker represents the 25th percentile minus 1.5 times the IQR. Circles represent statistical outliers. (Boxplot statistics calculated in R.)



Supplemental Figure 2.1. Computational modeling of CA XII active site. (A) Wild-type active site showing zinc ion (teal) coordinated by three histidines (119, 121, 145) and hydroxide group (red sphere) stabilized by threonine 226. **(B)** Mutation of amino acid 121 to glutamine leading to transient coordination of zinc by hydroxide ion and an increased theoretical bond distance of 3.04Å when compared to native histidine. Bond lengths and zinc distance shown by red and blue lines respectively.

	WT	Ex 9 Skip	Ex 9 and 10 skip	Ex 10 skip	Total Depth	% WT	% Ex 9 skip	% Ex 9 & 10 skip	% Ex 10 skip
Kidney 1	592	0	0	0	592	100.00%	0.00%	0.00%	0.00%
Kidney 2	454	0	2	0	456	99.56%	0.00%	0.44%	0.00%
Bronchial Epithelia 3	196	0	13	0	209	93.78%	0.00%	6.22%	0.00%
Skin 1	166	23	11	8	208	79.81%	11.06%	5.29%	3.85%
Fallopian Tube	27	153	2	0	182	14.84%	84.07%	1.10%	0.00%
Bronchial Epithelia 2	153	0	12	0	165	92.73%	0.00%	7.27%	0.00%
Colon 1	111	9	1	0	121	91.74%	7.44%	0.83%	0.00%
Panc 2	92	0	0	0	92	100.00%	0.00%	0.00%	0.00%
Skin 2	45	0	4	0	49	91.84%	0.00%	8.16%	0.00%
Nasal Epithelia 1	42	0	1	0	43	97.67%	0.00%	2.33%	0.00%
Colon 2	36	0	0	0	36	100.00%	0.00%	0.00%	0.00%
Thyroid	32	0	0	1	33	96.97%	0.00%	0.00%	3.03%
Nasal Epithelia 2	32	0	0	0	32	100.00%	0.00%	0.00%	0.00%
Panc 1	27	0	0	0	27	100.00%	0.00%	0.00%	0.00%
Small Intestine	24	0	0	0	24	100.00%	0.00%	0.00%	0.00%
Brain 1	0	12	0	0	12	0.00%	100.00%	0.00%	0.00%
Stomach	11	0	0	0	11	100.00%	0.00%	0.00%	0.00%
Cervix	2	5	0	0	7	28.57%	71.43%	0.00%	0.00%
Testes 1	3	3	0	0	6	50.00%	50.00%	0.00%	0.00%
Brain 2	0	5	0	0	5	0.00%	100.00%	0.00%	0.00%
Lung 1	2	0	0	1	3	66.67%	0.00%	0.00%	33.33%
Ovary	0	2	0	0	2	0.00%	100.00%	0.00%	0.00%
Bronchial Epithelia 1	1	1	0	0	2	50.00%	50.00%	0.00%	0.00%
Heart 1	0	1	0	0	1	0.00%	100.00%	0.00%	0.00%
Lymph Node	0	1	0	0	1	0.00%	100.00%	0.00%	0.00%
Lung 2	1	0	0	0	1	100.00%	0.00%	0.00%	0.00%
Testes 2	1	0	0	0	1	100.00%	0.00%	0.00%	0.00%

Supplemental Table 2.1. RNA-sequencing splice junction data shows distribution of CA XII isoforms in various tissues. Raw reads obtained from various tissues through the sequence read archive were aligned to the hg19 reference genome and splice junctions were detected using TopHat software. Depth indicates number of reads supporting each splice junction, with lower depths indicative of limited expression of *CA12* in that tissue. Percentages relative to all splice junctions detected for this region are

also included. All general tissue samples (non-epithelia) were obtained from the Human Protein Atlas through the SRA (ERP003613). Non-epithelia tissues are presumed to be of mixed cell types. Epithelia specific data were obtained from SRA studies: SRP018883, SRP044906, and SRP058237.

Chapter 3

Deep resequencing of *CFTR* in 762 F508del homozygotes reveals clusters of non-coding variants associated with variation in sweat chloride concentration and lung function

Briana Vecchio-Pagán, Karen S Raraigh, Rhonda G Pace, Matthew J Pellicore, Arianna Franca, Melissa Lee, Neeraj Sharma, Scott Blackman, Michael R Knowles, Garry R Cutting. *Submitted.*

Abstract

Extensive phenotypic variability is observed in cystic fibrosis (CF) patients homozygous for the common CF-causing variant F508del. To determine whether variants within and surrounding *CFTR* contribute to the phenotypic variability, we examined this locus in 1524 F508del chromosomes using next generation sequencing. In phase 1, ~200kb encompassing *CFTR* and extending 10 kb 5' and 5 kb 3' of the gene was sequenced in 486 F508del homozygotes selected from the extremes of sweat chloride concentration. In phase 2, a 510 kb region which included the entire topologically associated domain (TAD) of *CFTR* was sequenced in 276 F508del homozygotes drawn from extremes of lung function. An additional 163 individuals who carried F508del and a different CF-causing variant were sequenced to inform haplotype construction. Region-based burden testing of both common and rare variants revealed seven regions of significance (α=.01), five of which overlapped known regulatory elements or chromatin interactions. Notably, the -80kb locus known to interact with the *CFTR* promotor associated with variation in sweat chloride and lung function. Haplotype analysis revealed a single rare recombination event (1.9% frequency) in intron 15 of *CFTR* bearing the F508del variant. Otherwise, the majority of F508del chromosomes were markedly similar, with only minor variations of one ancestral haplotype observed, consistent with a single origin of the F508del allele. Together, these high resolution variant analyses of the *CFTR* locus suggest a role for regulatory motifs in trait variation among individuals carrying the common CF allele.

Acknowledgments

We would like to acknowledge David Mohr and CIDR for their thoughtful and methodical design of the resequencing capture. We would also like to thank the Cystic Fibrosis Foundation (CUTTIN13A2) and the National Institutes of Health (DK44003) for funding this work.

Introduction

Cystic fibrosis (CF) is a Mendelian disease affecting >70,000 patients world-wide. The primary phenotypes exhibited in affected individuals include elevated sweat chloride levels, lung disease, and pancreatic insufficiency. The severity of these phenotypes in each individual is primarily determined by the CF-causing variants they have inherited (*36*). CF variants can be classified into several groups by their expected functional impact on the cystic fibrosis transmembrane conductance regulator, or *CFTR* (*37, 126*). The most common disease-causing variant leads to the loss of phenylalanine at amino acid position 508, commonly referred to as F508del (ΔF508, rs121909001). This variant is present in the homozygous state in ~50% and heterozygous state in a further 40% of CF patients (*127, 128*), where it results in the improper folding and eventual degradation of the final protein product (*129*). In the absence of the CFTR protein, chloride and bicarbonate movement across the apical membranes of epithelial tissues is abnormal (*130*). Aberrant ion transport leads to unusually viscous secretions in the lung and pancreas, resulting in damage to both organ systems. While pancreatic enzyme replacement therapy counters the loss of pancreatic exocrine function, progressive

obstructive pulmonary disease is difficult to treat symptomatically and remains the primary cause of morbidity and mortality in CF (*131*).

Treatment of CF has been revolutionized by advent of small molecule drugs that target defective CFTR (*132, 133*). Due to the commonness of F508del, there has been intense effort to develop drugs that recover function of CFTR bearing F508del (*134*). However, F508del has been challenging to target as the omission of a single phenylalanine at codon 508 causes multiple defects in the function of CFTR. Prime among these is a folding defect that prevents CFTR from achieving a stable tertiary structure, leading to misfolded intermediates that are degraded by ER quality control mechanisms (*135-137*). Even when a small fraction of CFTR-F508del folds and traffics to the cell membrane, the resulting protein is minimally active (*138, 139*). These two defects have been tackled by combining a 'corrector' compound VX-809 that rectifies folding and a 'potentiator', VX-770 that activates the chloride channel of CFTR. The strategy has resulted in the recent approval in the U.S. of a combination drug (Orkambi™) for individuals homozygous for the F508del variant (*132, 140-142*). However, Orkambi produces only modest improvement in lung function in F508del homozygotes and no significant improvement in patient that have only one copy of F508del (*132*). Consequently, there has been a concerted effort to identify molecules with higher effectiveness for CFTR-F508del by empirical screening (*143*).

To inform the search for small molecules that target CFTR-F508del, we have systematically investigated genetic variation in *CFTR* alleles bearing F508del. Our overarching goal was to identify any variants or combination of variants that modify disease severity associated with the F508del mutation. Indeed, examples of common

(R117H and poly T tract; (*144*)) and rare (F508del and R553Q; (*145*)) intragenic modifiers in *CFTR* have previously been reported. Studies of F508del homozygotes have revealed a broad range of both sweat chloride values and lung function (*36, 44, 45, 146*). Affected twin and sibling studies indicate that the *CFTR* locus is the primary determinant of variation in sweat chloride concentration, accounting for 56% of total variance (Collaco et al, *in review*). Variation elsewhere in the genome (i.e. genetic modifiers) appears to play a minor role in sweat variability. Variation in the *CFTR* gene also contributes to phenotypic variance in lung function but at a lower effect size (*46*). Thus, F508del homozygotes at the extremes of the distribution of CF traits, particularly sweat chloride concentration, provide an opportunity to find variants that modulate the effect of F508del upon CFTR function.

Identification of variants *in cis* with F508del that modify disease severity should improve our understanding of the molecular mechanisms that ameliorate the defect caused by the F508 variant. These findings could, in turn, inform rational design of molecular treatments for CF. Consequently we have performed deep re-sequencing of intragenic and extragenic regions surrounding *CFTR* in 762 F508del homozygotes. Region-based burden testing was used to test whether genetic variation within and near *CFTR* which is associated with sweat chloride concentration or lung function. High resolution haplotypes and linkage disequilibrium patterns were used to map recombination events on the F508del bearing chromosomes.

Results

Variation in *CFTR* chromosomes bearing F508del

A total of 925 individuals with CF (762 F508del homozygotes and 163 F508del heterozygotes) were sequenced in two phases. To increase power to detect associations with modifying variants in the *CFTR* locus, all available F508del homozygous individuals in the Johns Hopkins CF Twin and Sibling Study (TSS) and the Genetic Modifier Study (GMS) with extremes of sweat chloride levels (**Figure 3.1A**) were selected for analysis. A 210kb region encompassing *CFTR* and extending 10kb 5' and 5kb 3' of the gene was sequenced in 583 subjects (486 F508del homozygotes and 97 heterozygotes: **Table 3.1)**. Advances in capture technology following completion of the first phase enabled us to expand coverage of the *CFTR* locus in a second phase. Consequently, in phase 2, a 305 kb region fully encompassing the topologically associated domain (TAD) of *CFTR* (*147, 148*) plus an additional 300kb flanking this TAD was sequenced in 342 subjects (276 F508del homozygotes and 66 heterozygotes). The expanded region includes the neighboring genes *WNT2*, *ASZ1*, and *CTTNBP2*. In the second phase, we selected the F508del homozygotes in the TSS from the extremes of lung function (**Figure 3.1B)**. By combining the two phases, we obtained F508del homozygous individuals drawn from the entire phenotype spectrum for sweat chloride function and lung function (**Figure 3.2**). The F508del heterozygous samples were used only to inform haplotype studies. Otherwise, the following results are specific to the sequence-verified F508del homozygous population (n=762).

A total of 652 variants (both SNPs and INDELs) were observed within *CFTR* in the F508del homozygous subjects. Twenty four variants were observed within *CFTR* exons (n=13) **(Tables 3.1, 3.2)** or in the 5' and 3' UTRs (n=11) while the remaining 628 were intronic. Two of the 13 exonic variants had MAF > 1%, (p.I1027T and p.Q1463Q), while a third variant, L467F (rs1800089), was detected at a frequency of 0.3% in *CFTR* genes bearing F508del. Of note from a clinical diagnostic perspective, L467F has been reported in individuals also carrying 1 copy of F508del diagnosed with CF related metabolic disorder (*149-152*). Our results indicate that L467F is *in cis* with the F508del allele, indicating that it is not the *trans* CF variant leading to disease in these patients. However, due to low frequency in this population, we were not able to resolve the revertant potential of L467F or three other variants that cause an amino acid substitution (p.Q1330Q, p.R1438Y and p.V1475M) or five synonymous variants (p.L130L, p.I203I, p.T854T, p.I1404I and p.Q1463Q). None of these variants was predicted to activate cryptic RNA splicing (M. Lee, novel splice prediction algorithm, manuscript in preparation). Functional testing will be required to assess whether the amino acid substitutions affect the function of *CFTR* bearing F508del. Of note, the p.V1475M allele detected here was present in one of the original *CFTR* cDNAs that was widely distributed, and it appears to have no functional effect (*153*). Finally, five variants initially mapped to the *CFTR* gene were discovered to be variants in regions of high homology to *CFTR*'s exon 10 (*154, 155*) **(Table 3.3).** Together, these results indicate that there is limited variation in the coding regions of *CFTR* bearing the F508del mutation.

Single variant association analysis reveals no significant association with sweat chloride concentration or lung function

Common variants with MAF > 1% (602 total variants; 235 in phase 1; 598 in phase 2) were assayed for association with either sweat chloride levels **(Table 3.4)** or lung function as measured by the age and survival adjusted phenotype SAKNORM **(Table 3.5)**. Linear regressions between number of minor alleles and the phenotypes were conducted on the Phase 1 and Phase 2 study samples separately and, for regions sequenced in both captures (the 210kb encompassing *CFTR*), in combination. All P-values were calculated by permutation because of possibly non-normally distributed phenotypes (Supp Fig. 1C). 43 variants showed some evidence of association with point-wise permutation p-values ($p<.05$) and 15 variants had beta values in the same direction when observed in both phase 1 and phase 2 cohorts (**Tables 3.4 and 3.5**). None of these variants was statistically significant after multiple test correction using either max(T) permutation (Supp Table 3) or by Bonferroni correction (not shown).

One coding variant (p. I1027T) showed weak evidence of association with lung function (uncorrected point-wise permutation P=.0130). Recognizing the limited power to detect associations of single rare variants with our traits (57% and 18% power at MAF .01 and an effect size of 1SD for sweat chloride and lung function respectively), we tested I1027T for association in a second, unrelated group of 748 F508del homozygotes. However, the association of I1027T with lung function did not replicate (P=0.8239). These results indicate no single common variant *in cis* with F508del was associated with either sweat chloride concentration or lung function in this study.

Clusters of variants 5' of and within *CFTR* correlate with sweat chloride concentration and lung function in F508del homozygotes

A region-based burden assay was performed to identify groupings of variants which display associated with trait variation. Variants were tested for association with sweat chloride or lung function in groups defined by a series of overlapping 5kb windows (offset in 1250bp increments) with each window generating a test P-value. The P-value for each window was Bonferroni corrected for multiple testing based on the total number of unique windows assayed (see Methods). Regions of significance (hg19 coordinates, study-wide P<0.01) were highlighted for the combined test of common and rare variants associating with either sweat chloride concentration (**Figure 3.3)** and/or lung function (**Figure 3.4**). Regions that coincided with known regulatory and boundary elements are discussed below (**Table 3.6**) (*148, 156, 157*).

A regulatory locus at -80kb is associated with both sweat chloride levels and lung function

Variation in a locus denoted 'Region A' was associated with both sweat chloride concentration (study-wide corrected P=5.9e-4; region of statistical significance: chr7:117,039,250-117,053,000, **Figure 3.3**) and with lung function (study-wide corrected P=8.9e-3; chr7:117,030,500-117,050,000, **Figure 3.4**). Region A is located approximately 80 kb 5' of the *CFTR* transcription start site and within intron 4 of *ASZ1*. Chromatin conformation capture assays have shown that this region interacts with sequences in the *CFTR* promotor (*148, 156*). It also contains CTCF binding sites (**Figure 3.3**), which may assist in looping of distant regulatory elements to *CFTR*'s transcriptional

start site (*148, 157, 158*). A total of 12 common and 9 rare variants are observed under the sweat chloride peak (**Table 3.7**). Of note are two variations in the length of a poly A tract, which may both be associated with higher sweat chloride concentration (7:117047463:TA>T and 7:117047463:TAA>T. A total of 13 common and 3 rare variants are observed under the lung function peak **(Table 3.8)**. One of these variants (rs4730780, 7:117041448:T:A) that results in a decrease in length of a poly T tract, and an increase in length of the adjacent poly A tract, is located ~100bp from a known CTCF binding site(*159*). The 4 individuals with this variant all had above average lung function (beta = +0.81 SAKNORM).

Three Loci Associated with Sweat Chloride Levels

Variation in 3 regions associated with sweat chloride concentration (**Figure 3.3**). Region B (chr7:116,941,750-116,951,750, P=7.1e-3, phase 2 coverage only, n=276) is located within intron 3 of *WNT2* and contains 21 common and 9 rare variants (**Table 3.7**). This region lies outside *CFTR*'s topologically associated domain (TAD), and is adjacent to known CTCF binding sites in MCF7 and K562 cells (**Figure 3.3**). Three individuals harbored what appears to be a haplotype of 4 rare variants (116942115:G>A, 116942433:G>T, 116943135:C>T, 116944283:T>A, 3/552 chromosomes) and had lower mean sweat chloride concentration (~17mM Cl-). These variants lie within or near a region of open chromatin and CEBPB binding site in fetal lung fibroblasts cells (IMR90) (*159*). A common variant in the same region also showed evidence of association (P=0.02 uncorrected) with decreased sweat chloride levels (7:116943793:A>T, -8.63 mM Cl-). Region C (chr7:117074250-117078000, P=3.3e-3, phase 1 and phase 2 coverage, n=762) contains 6 rare variants and a known regulatory locus ~44kb upstream of

CFTR(*148*). The variant with the most significant uncorrected p-value (117076029:G>A, P=0.001648, **Table 3.7**) lies within MTA3 and PML binding regions in GM12878 cells (*160*). The final region significantly enriched for variants associating with sweat chloride levels (Region D, chr7:117153000-117156750, P=5.8e-3, phase 1 and phase 2 coverage, n=762) is located within intron 3 of *CFTR,* and contains 2 common and 7 rare variants. Region D contains no known functional elements and has various repetitive sequences.

Two Distinct Loci Associated with Lung Function

Variation in 2 regions associated with SAKNORM (**Figure 3.4**). Region E (chr7:117,010,500-117,014,250, P=.0058, phase 2 coverage only, n=276, intron 10 *ASZ1)* lies just outside of the proposed TAD containing *CFTR* **(Figure 3.4)**, and contains 6 common and 1 rare variants (**Table 3.8**). Most of the variants are 2-4kb 3' of known CTCF binding sites, but may influence interactions with regions outside of the *CFTR* TAD. Region F (chr7:117159250-117164250, P=3.3e-3, n=762) is located within intron 3 of *CFTR,* slightly 3' to region B associated with sweat chloride **(Figure 3.3)**. Association here appears primarily due to common variation, specifically variations within a poly T tract (7: 117160319, 17T, 18T, or 16T). Increasing the length of this tract is marginally associated with improved lung function (uncorrected P=0.0036), and conversely, decreasing the length of this tract is associated with poorer lung function (uncorrected P=0.0033) (**Tables 3.4, 3.5, and 3.8**). The region has a large number of repetitive elements and no known functional elements.

One recombination event defines two blocks of linkage disequilibrium within *CFTR* bearing F508del

In addition to association testing, we also sought to systematically determine the genetic architecture at this locus in the F508del homozygous population. Haplotypes are combinations of variants that tend to be inherited together (*in cis*). Their borders are often delineated by recombination events which occur during meiosis. Derivation of haplotypes are useful for establishing the degree of genetic diversity in a locus, associating functional variants with background variation and inferring ancestral origins of disease-causing variants. To assemble haplotypes, single nucleotide variants with MAF > 1% within the 510 kb region surrounding *CFTR* were phased using SHAPEIT2 (*161*). Samples bearing non-F508del chromosomes were used to deduce locations of alternative recombination events and additional diversity of the *CFTR* locus. Linkage disequilibrium (LD) among variants with MAF > 2% in 762 F508del homozygotes revealed three primary regions of high LD, two of which encompassed *CFTR* **(Figure 3.5).** A recombination event was observed within intron 15 of *CFTR*, resulting in an alternative haplotype in LD block 2. This recombination event is not present in 206 non-F508del chromosomes which were sequenced using the same capture design. The intron 15 recombination event in F508del homozygotes is unique to this population, and is not apparent in HapMap populations (which contain a diversity of *CFTR* haplotypes only ~5% of which is the F508del ancestral haplotype), where a distinct recombination event within intron 22 is observed (*162*).

Using common variants with a MAF > 1%, we found that 16 haplotypes occurred at a frequency of 0.5% or higher in LD block 1 (*CFTR* exons 1-15) on chromosomes bearing F508del **(Figure 3.6, Table 3.9).** The second LD block encompasses exons 16-27 of *CFTR*, and 10 haplotypes above 0.5% frequency are observed in the F508del

homozygous cohort. Additional diversity is seen via inclusion of INDELs, but due to the large number of such sites observed in this region and their poly-allelic state, they were not included in the analysis presented.

LD block 1 of *CFTR* encompassing the F508del allele displays primarily rare variation

Limited common variation would be expected in the region of high linkage disequilibrium surround the F508del variant if it were inherited from a common founder ancestor of European descent (*127*). Indeed, the autozygosity of this region is confirmed by the dearth of common variants in this haplotype (2/17 SNPs with MAF > 5%). The majority of variation within this locus is very low frequency (15/17 SNPs with MAF between 1-5%). Overall, ~70% of F508del homozygotes carry the same ancestral haplotype in LD block 1, with minor variations of this haplotype occurring in another 30% of samples **(Figure 3.6)**. The majority of haplotypes in LD block 2 deviate from the ancestral haplotype by only one marker. One such haplotype contains I1027T (rs1800112), a variant known to be observed *in cis* with F508del (*163*). Our data, as well as data from the *CFTR*2 database (K. Raraigh and P. Sosnay, personal correspondence), indicate this allele and its associated haplotype are present on ~2.5% of F508del chromosomes **(Figure 3.6,** variant #1239, red asterisk).

LD block 2 of *CFTR* contains two distinct haplotypes

Due to the historical recombination event in intron 15 which occurred on an F508del-containing chromosome, a second distinct haplotype is observed in LD block 2 that represents 1.9% of F508del chromosomes **(Figure 3.6,** bottom right). The alternative haplotype is observed in two forms (0.6% and 1.3% MHF), which vary by one

marker (rs138427389) **(Figure 3.6,** variant #1305**)**. Of note, this alternative haplotype contains the synonymous variant Q1463Q (rs1800136) **(Figure 3.6**, variant #1507). When samples bearing this alternative haplotype are removed from the F508del homozygous population, the recombination event within intron 15 is not observed (**Figure 3.7**). The majority of variation observed in LD block 2 is rare (46/47 SNPs with MAF between 1-5%). Overall, 85% of F508del chromosomes carry the ancestral haplotype in LD block 2, with the alternative haplotype as well as minor variations representing another 15% of chromosomes. Finally, when considering only SNPs with MAF > 1% within the *CFTR* locus, ~55% of F508del chromosomes are completely identical across both regions of LD.

Thirteen single nucleotide variants capture all common (>1% frequency) haplotypes found on *CFTR* chromosomes bearing F508del

By definition, haplotypes of a given LD block contain SNPs in high linkage disequilibrium. As such, these haplotypes can be simplified by tagging variants. In this process, all variants above a particular LD threshold are grouped and represented by a single variant, or 'tag'.

All haplotypes composed of SNPs with a MAF > 1% within the ~200kb surrounding *CFTR* are detailed in supplemental table 7, where a subset of 31 tagged SNPs **(Table 3.10)** captured 99.2% of the total 'common' variation at an r^2 correlation value greater than 0.9. Of these tagged SNPs, 13 were capable of representing all haplotypes with a minor haplotype frequency (MHF) > 1% (**Table 3.9,** grey highlights).

Haplotype-based studies using 13 tagged SNPs did not reveal any significant association with either sweat chloride concentration or lung function (data not shown).

Mapping of Restriction Fragment Length Polymorphisms at the *CFTR* Locus

CFTR was originally mapped using restriction fragment length polymorphisms (RFLPs) (*32, 164*). We mapped 8 of these RFLPs to within 2kb of their hg19 genomic coordinates, and have determined the coordinate and rsID of an additional 3 RFLPs (**Figure 3.8**). These RFLPs primarily lie within LD block 1 (Intron 3 *WNT2* to Intron 15 *CFTR*), which includes the F508del allele. Even RFLPs lying beyond this linkage block show residual LD with the F508del allele. An example of this is H2.3A (XV-2C, rs3779549), which lies just beyond the recombination event in *WNT2*, but displays residual LD with the F508del locus. This allele likely was on the same haplotype as F508del, but over time, recombination events decreased LD between these two markers. For this reason, the reference allele of H2.3A is enriched in the F508del population. As the majority of CF patients carry at least one copy of the F508del allele, the high LD of these markers with this variant facilitated the localization of *CFTR* (*34*).

Discussion

We have systematically characterized both common and rare variation within and surrounding the *CFTR* locus in a large cohort of F508del homozygotes. Our initial goal was to determine if a coding region variant might moderate the deleterious effect of the F508del allele. A small number of rare exonic variants were observed *in cis* with F508del. We found that neither a single common variant nor a combination of variants (i.e. haplotype) within this region is associated with CF trait variation. However, we

were unable to exclude four rare amino acid substitutions (p.L467F, p.Q1330E, p.R1438Y, and p.V1475M) as these variants were not frequent enough to allow for statistically valid association testing in this population. Functional studies will be required to assess if these rare variants have any effect on *CFTR* bearing the F508del allele. By extending analysis to non-coding sequences, we identified clusters of variants associating with variation in sweat chloride levels and/or lung function in proximity to regulatory elements 5' of *CFTR*. Using the rich dataset produced by sequencing the entire *CFTR* locus, we were able to resolve the genetic architecture surrounding the common CF-causing variant to an unprecedented level of detail. Haplotype analysis revealed that a single recombination event in intron 15 occurred in *CFTR* bearing the F508del variant. This event generated two regions of high linkage disequilibrium (LD) encompassing the *CFTR* locus. The LD block containing F508del extends into the *WNT2* gene, thereby facilitating the mapping of the *CFTR* locus. The dearth of common variation on haplotypes bearing the F508del variant is consistent with its single ancestral origin (*163*). Finally, we were able to extract 13 haplotype-tagging SNPs that represent the vast majority of the genetic variation surrounding F508del. The SNPs could be used to parse F508del homozygotes into sub-populations to test whether variation at the *CFTR* locus underlies differences in responses to molecular-targeted treatments.

Basal transcription of *CFTR* is primarily driven by binding of factors at the 5' promotor element (*165*). However, recent studies have shown that additional *cis* regulatory elements are required for tissue specificity, abundance, and temporal expression (*166, 167*). These *cis* regulatory elements have been shown to interact with the *CFTR* promotor, likely through a chromatin looping mechanism in part facilitated by

CTCF binding (*168*). Multiple chromatin interaction studies have now shown that these regulatory loci are encompassed within a topologically associated domain or TAD (*148, 157, 169*), which is defined by boundary elements at -80kb and +49kb from *CFTR*. While much progress has been made regarding the chromatin structure in this region, resolving the function of each of these regulatory elements continues to be an active area of research.

The results presented here posit that a burden of both rare and common variants at key loci may modulate the CF phenotype by alteration in the level and/or timing of expression of *CFTR* bearing F508del. These findings may inform future functional studies of the *cis*-regulatory elements identified in chromatin studies. The shared burden of rare and common variants associating with both CF traits at the -80kb regulatory motif is possibly the most striking finding we report. We hypothesize that variants here may affect CTCF binding, or increase inherent enhancer activity. This could lead to altered expression of the F508del transcript, which has some residual processing and function (*96, 170-174*). Presence of even small amounts of partially functional CFTR over the lifetime of an individual might be sufficient to moderate CF traits such as sweat chloride concentration and lung function (*171*). The concept that natural variation in the expression level of mutated genes may underlie differences in the severity of inherited diseases is supported by recent studies of loss-of function yeast phenotypes (Vu et al Cell 2016). Additionally, we posit that the intragenic and extragenic variation present in the F508del population may confer increased or decreased response to Orkambi™ or future CFTR specific drugs.

The most 5' regions of interest (Regions B and E, **Figures 3.3 and 3.4**) were located within introns of *WNT2* and *ASZ1*. The region in *WNT2* is located in an adjacent TAD to *CFTR*. In a recent study, there was no report of this region interacting with the *CFTR* locus (*157*). However, previous studies in epididymis cells indicate there are weak long range chromatin interactions with this region that may be cell type specific(*148*). It is possible that some of the rare variants associating with sweat chloride in the region are somehow modifying overall chromatin organization in certain cell types, such as the sweat gland. Another distant region of interest was found within *ASZ1*, and is located just outside of the proposed *CFTR* TAD(*157*). These regions are often enriched for TAD-TAD interactions, also known as inter-TAD interactions. Variants here could alter CTCF binding, TAD architecture, or inter-TAD interactions. Assaying both of these possible inter-TAD interactions could lead to additional insight into distant regulatory elements in *ASZ1* and *WNT2*. Of note, the 5' TAD boundary proposed by Smith and Dekker (2016) closely follows the recombination event in intron 10 of *ASZ1* in this study, suggesting a possible link between recombinant events and chromatin structure in this region.

Interestingly, adjacent regions within intron 3 of *CFTR* were found to associate with both sweat chloride levels and lung function (albeit, there is no distinct overlap given the coordinates identified here). To our knowledge, this region has not previously been shown to have regulatory function. While the intron 3 signal for sweat chloride was primarily composed of rare variation, common variation in the length of a poly T tract resulted in the lung function association. Interestingly, while not significant, the 18T and 16T alleles at this locus also showed some signal in sweat chloride levels (P=.06, beta= -

3.58 mM Cl- and P=.08, beta=+3.49 mM Cl- respectively). Given the lack of functional elements and low conservation in this region, it is currently challenging to imagine a mechanism by which this alteration could modulate CF traits. However, poly T tracts may regulate gene expression by acting as matrix attachment regions (MARs) (*175*), or may participate in RNA triplex formation (*176*). We do note that this poly T tract still may modulate lung function when the cohorts are considered independently (18T allele: P=.051, beta=+0.18, n=486 and P=0.12, beta=+0.19, n=276).

A recurring theme through-out the variation observed in this study was variable lengths of repetitive elements associating with disease severity. These INDELs may represent a mode of phenotype modification that is not well characterized (*177*), but has been previously observed to modify the phenotype of other CF-causing alleles (i.e. R117H and polyT tract) (*144*). This type of variation is observed in 5 of the 7 regions of interest. We recognize that this form of variation can be difficult to accurately call using next generation sequencing methods. However, all INDELs reported here were of both high mapping and variant call qualities, and exhibited Hardy Weinberg equilibrium. It is possible that these small variants may be partially marking larger repetitive sequences that could not be typed in this study due to read length (or high homology). Additional studies of common and rare INDELs at these loci could reveal a mechanism of phenotype modification. Finally, we recognize that some of the assays employed here have limited power, especially at low minor allele frequencies given the cohort size (which ranges from 276 to 762, depending on the region assayed). Power is additionally limited when assaying sweat chloride associations in the phase 2 cohort, as this cohort was selected for extremes of lung function, and thus contains intermediate sweat chloride values. Given

these limitations, the study presented here likely contains false negatives, which could only be resolved using larger cohorts.

Some sequencing studies fail to consider regions of known homology with the region of interest. In this study, we opted to allow for a higher frequency of false positives in regions of the capture with high homology to pseudogenes (specifically intron 9 and exon 10 of *CFTR*, **Table 3.3**) (*155*). This was to allow for more consistent tiling of baits, better detection of large structural variants, and a more complete capture overall. Clinical labs should be aware of these regions when designing assays in order to minimize erroneous calls. For example, the nonsynonymous mutation A455E is a high frequency CF-causing allele in exon 10. This variant is also present in a pseudogene present on chromosome 20. While this variant can be correctly typed using a longer read length, short amplifications cannot distinguish between these two forms (*178*). The variants reported in **Table 3.3** could be assigned to either the chromosome 9 or chromosome 20 pseudogenes due to their reoccurrence in a small subset of samples (n=5); however, alternative methods would be required in a clinical setting.

Overall variation was rare within the LD block containing F508del, consistent with a single ancestral origin of this allele in the population. When considering only common SNPs (MAF>1%), the majority of F508del chromosomes (~55%) are completely identical. These results indicate the F508del homozygous population is highly homogenous, with the majority of variation being private or due to a low-frequency recombination event within intron 15. Because this event is not observed on non-F508del chromosomes, it likely occurred after F508del arose. Previously, a recombination event was reported to have occurred within intron 22 (*162*). However, this recombination event

was based on population level data provided by the HapMap project, which used wild-type *CFTR* and had significantly reduced marker density compared to this study (*179*). Newer HapMap releases suggest two possible primary recombination events in the general population: intron 11 and intron 15-intron 16. The intron 15-intron 16 event likely represents F508del carriers, as it is primarily observed in European populations. The previously reported intron 22 event may have some limited evidence in Mexican and Italian Hapmap cohorts.

We have now made available a detailed map of common variation and population based haplotypes for the F508del locus **(Table 3.9)**. In summary, this study has methodically characterized variation *in cis* with the F508del allele and the genetic architecture of this locus in great depth. Collectively, our findings suggest a combination of rare and common variation within suspect and known regulatory regions at the *CFTR* locus may contribute to the phenotypic heterogeneity observed in F508del homozygous CF patients. The identified variation may modify *CFTR* expression levels and/or timing of expression, and should inform future regulatory studies of this locus.

Methods

Cohort Selection:

To increase the power to detect associations with CF traits, patient samples were selected at the extremes of phenotypes (*180*). These extremes were defined by plotting a normal distribution of these traits in F508del homozygotes which are part of the Johns Hopkins Twin and Sibling Study (TSS) (*181*) **(Figure 3.1)**. Samples selected for the phase 1 of this study had sweat chloride levels +/- 1SD from the population mean, and were recruited from either the TSS or the University of North Carolina's Genetic Modifiers Study (GMS). Of note, the distribution of sweat chloride in the GMS was highly similar to that of the TSS. In phase 2 of the study, samples were similarly selected from extremes of SAKNORM, an age and mortality adjusted CF-population specific measure of lung function ***(182)***, and were recruited from the TSS.

Sample consent and DNA preparations:

Patients in this study were recruited to either the Johns Hopkins Twin and Sibling Study(*181*) (TSS) or the University of North Carolina's Genetic Modifiers Study (*183*) (GMS). Peripheral blood was previously obtained from all consented individuals as part of their respective studies, and genomic DNA was extracted by a phenol-chloroform protocol. Phenotype information for TSS and GMS cohorts was collected as part of their respective studies, and was taken from both study admission forms, as well as the Cystic Fibrosis Foundation's Annualized Database.

Re-sequencing Capture Design:

Capture design and sequencing was conducted at the Center for Inherited Disease Research (CIDR), which implemented a 3-tiered process to design the reagent. The first pass includes 1x tiling density with 'most stringent' masking of repetitive regions, and maximum performance boosting for probe re-balancing which replicates any GC rich probes based on Agilent's exome capture algorithm to improve uniformity. Regions not covered from the first pass were re-submitted with relaxed parameters (1x tiling, moderate stringency masking and balanced boosting) in an attempt to provide complete coverage across the regions of interest. Any regions not covered from this second pass were re-submitted using a lower stringency (1x tiling, least stringency masking and no boosting). All probes obtained at each pass were analyzed using BLAT, allowing determination of probe uniqueness, which aided in estimating impact on percent selection and mapping quality. Any probe which had a BLAT score of <40 was flagged, and flagged probes were reviewed with the authors for further evaluation.

Library Prep, Capture, and Sequencing:

1ug of genomic DNA was sheared using the Covaris E210 instrument (Covaris), shear time was decreased to 80 seconds in order to obtain larger insert sizes. A hybrid protocol for library preparation and whole exome enrichment was developed at CIDR(unpublished) based on methods and parameters from Fisher, applied to the reagents, volumes and parameters from the Agilent SureSelect XT kit and automated protocol (p/n G7550-90000 revision B). All processing was done in 96 well plate formats using robotics (Beckman F/X, Perkin-Elmer Multiprobe II, Agilent Bravo, Beckman Biomek 2000). 'With Bead' clean ups were used following shearing, end repair, A-tailing and adapter ligation. The initial input of Ampure XP SPRI (Beckman

Coulter Genomics) beads was based on volumes from the Agilent protocol. After the first clean up the sample was eluted and the beads remained in the reactions through the final ligation clean-up. These reactions were carried out using the XT reagent kits, volumes and conditions described in the Agilent protocol. At pre-capture PCR the entire product was amplified, adjusting the water in the reaction to accommodate the increase in DNA sample volume. The PCR enzyme used in all steps was switched from Herculase to Kapa Biosystems HiFi HotStart Ready Mix. The Kapa enzyme increases coverage in GC rich regions. All other aspects follow the Agilent protocol except the number of PCR cycles was increased from 6-8 cycles. 750ng of amplified library was used in an enrichment reaction following Agilent protocols (24 hour hybridization). Post-capture washing was done using the Agilent protocol except the 'off-bead' catch process from Fisher et al., was incorporated (samples are not eluted off the DynaBeads (Invitrogen), instead post-capture PCR master mix and indexes are added directly to the beads). Post-capture PCR was done according to the Agilent protocol, with the adjustment of water volume and PCR cycles where needed. For low input samples, 50ng of genomic DNA was sheared using the Covaris E210 instrument (Covaris), using the same parameters as the 1ug input samples. Library prep was performed using the Kapa Hyper Prep kit (Kapa Biosystems) according to the manufacturer's protocol. Indexed adapters, primers and blockers used in the low input protocol were custom synthesized by Integrated DNA Technologies and were used in accordance to the Kapa protocol specifications. The hybridization for capture was set up according to the Agilent XT protocol with the exception that blocker #3 was exchanged for the IDT custom synthesized blocker. Post-

capture washing was done using the Agilent protocol except the 'off-bead' catch process from Fisher et al., was incorporated (samples are not eluted off the DynaBeads (Invitrogen), instead post-capture PCR master mix are added directly to the beads). Post-capture PCR was done according to the same parameters used for the pre-capture PCR adjusting for PCR cycles. DNA sequencing was performed on an Illumina® HiSeq 2500 instrument using standard protocols for a 100 bp paired-end run. 96 samples were run per flowcell, guaranteeing >90% completeness at a minimum of 20X coverage. Intensity analysis and base calling were performed through the Illumina Real Time Analysis (RTA) software (version 1.17.20). Basecall files were converted from a binary format (BCL) to flat file format (qseq.txt) using the Illumina BCL Converter software (version 1.9.4). qseq.txt files were demultiplexed to single sample fastq files using a demultiplexer written at CIDR as part of CIDRSeqSuite version 3.2.

Variant Calling and Annotation:

Briefly, reads were first aligned to the genome using Burrows-Wheeler Aligner Mem algorithm (*115*) (v.0.7.8), duplicate reads were marked using PicardTools (v.1.109), realignment around INDELs and base quality score recalibration were performed via the Genome Analysis Tool Kit (v.3.1-1)(*83*). Small variant calling was conducted using three software: GATK's assembly based haplotype callers, FreeBayes, and Platypus. The final call-set required that a variant or INDEL be called by at least two algorithms (custom scripts allowed adjustment of INDEL call positions). In GATK, per sample gVCFs were used for joint calling of the entire cohort. The call-set was further refined by hard filters (VQSR not applicable in this targeted dataset). Variant filters generally included a depth of coverage >=10x, variant quality scores >=30, and read mapping

quality >=40. Alternative and additional filters were applied to INDELs. Large structural variants were called via Conifer and Breakdancer software. Custom scripts then annotated the call set for disease causing *CFTR* variants as defined by CFTR2, and typed known repeats of interest (poly T and TG tracts).

Data Cleaning:

The *CFTR* genotype of each of the samples in this cohort was checked for concordance to study admission records. Of the 790 F508del homozygotes sent for sequencing, 15 were found to carry alternative genotypes (i.e. G551D, I507del, etc). These samples were removed from the F508del cohort. Given that nearly all samples sequenced had SNP level genotypes available from a previous genome wide association study (GWAS), concordance was checked between the 82 typed SNPs in our targeted regions and the GWAS call set. The overall concordance rate was 99.2%. Discordant samples (likely plated erroneously in either study) were removed from the analysis. One sample was found to have Klinefelter syndrome (SNP data consistent with a karyotype of 47,XXY and verified via clinical contact), and was removed from the analysis. Two samples were found to contain large structural deletions and were removed from the analysis. Identity by state analysis was used to predict unknown related samples. Population stratification was assayed via principle component analysis. No significant stratification was observed beyond the second ancestral haplotype in LD block 2 of *CFTR* (principle component 1). This stratification did not correlate with either reported ancestry or known HapMap populations. The entire cohort most closely resembles TSI and CEU HapMap populations. The above analyses were conducted using combinations of Plink v1.9 (*184*), R, and custom scripts.

Common Variant Association Testing:

Variants with MAF > 1% in the F508del homozygous population were assayed for association with either sweat chloride levels or SAKNORM using linear regression analysis in the Plink software package v1.9(*184*). Data was initially cleaned for individual and SNP missingness, and IBD structure (to remove a select few related samples). Empiric point-wise and study-wide corrected P-values were calculated using Max(T) permutation testing (PLINK --mperm command with $1x10^6$ permutations).

Common and Rare Variant Burden Testing:

The optimized sequence kernel association test (SKAT-O) was used in this study as it maintains power when a large portion of variants are neutral or have opposing direction of effects (*185*) (power for 762 subjects is ~70% in regions where 20% of the variation is non-neutral). To increase power for this test, the combined cohorts were used to test association in regions of overlapping coverage (**Figure 1**, bottom track). Common and rare variants were grouped according to a MAF cut-off of 1% in the F508del homozygous population. A 5kb sliding window was then moved across the entire capture region in 1250bp increments. Common and rare variants falling within these regions were grouped for region based burden testing using the SKAT-O algorithm, implemented in R. Variants were assayed for either common, rare, or combined burden testing using the "SSD" commands which allow for loading of a plink formatted dataset. The method employed in all cases was "optimal.adj", representing the optimized method. The resulting region based burden p-values were tested for study-wide significance by correcting for the total number of unique SKAT-O tests performed. Duplicate windows

which contained the same set of variants as another window were not counted as independent tests. In the 510kb captured region encompassing the *CFTR* locus and surrounding genes, a total of 404 windows were present. Of these 320, 186, and 374 were independent tests for rare, common, and combined tests, respectively. All SKAT-O derived p-values were corrected for the above mentioned multiple testing. Statistical significance was defined as a P-value less than 0.01 after correction for the number of windows used in the analysis.

Phasing and Haplotype Analysis:

Variants in the F508del homozygous population (n=762) with MAF > 1% were phased using phase informative reads in SHAPEIT2, with recommended settings. LD blocks and haplotypes were confirmed and visualized using Haploview. To convey relevant information from supplemental Figure 1 in a non-matrix format, a custom perl script was used to calculate the ratio of r^2 correlation values on either side of a given variant. Plink v1.9 was first used to generate r^2 values for up to 20 possible variant pairs within 10kb of a given SNP (--r^2 --ld-window 20 –ld-window-kb 10 –ld-window-r^2 0). These values were entered into a script which divided these pairings into variants either upstream or downstream of the locus in question. The recombination ratio was then calculated as: [mean r^2 of the SNP with upstream variants / mean r^2 of the SNP with downstream variants]. The recombination ratio by distance was then plotted using R.

Haplotype Based Association Testing:

CFTR haplotypes as defined by tagged variants in **Supplemental Table 6** were assayed for association with either SAKNORM or sweat chloride levels in the F508del

homozygous cohort. This analysis was conducted in Plink v.1.07 using the --chap and --each-vs-others commands. Only haplotypes with frequencies > 1% were assayed.

Power Calculations:

Study phase specific power calculations were conducted using custom R scripts adapted from Guey and Voight 2011 (*186*). Key parameters include the sample size (n=486 for phase 1, n=276 for phase 2), extremity of phenotypes (+/- 1 SD), and the percentage of phenotypic variance attributable to mutations within the *CFTR* locus (50% for sweat chloride and 10% for SAKNORM (Collaco, *submitted*).

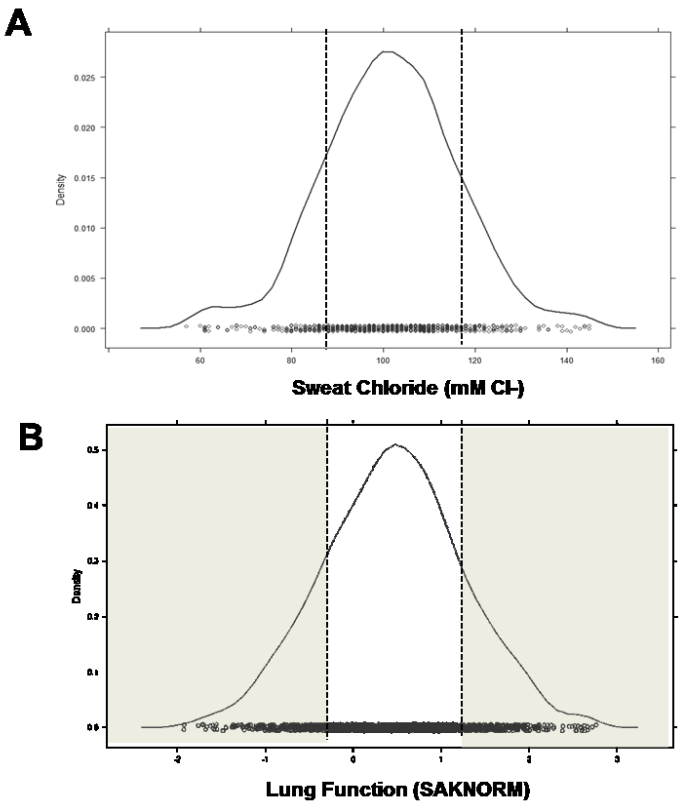


Figure 3.1. Selection of individuals from extremes of distribution of sweat chloride concentration or lung function.

A) Density plot of sweat chloride levels in F508del homozygotes in the Johns Hopkins Twin and Sibling Study (TSS; n =784). Dotted lines indicate thresholds used for selecting samples for phase 1 of this study at either -1 or +1 SD (~15mM Cl-) from the mean (~101.5mM Cl-). Using these thresholds, 231 samples were selected from TSS, and 255 from Genetic Modifier Study (GMS) for phase 1. B) Density plot of SAKNORM levels in F508del homozygotes in the Johns Hopkins Twin and Sibling Study (TSS; n=1205). Dotted lines indicate threshold for selecting samples for phase 2 of this study at either -1 or +1 SD (0.8 SAKNORM units) from the mean (0.45 SAKNORM). A total of 276 samples were selected from TSS for phase 2.

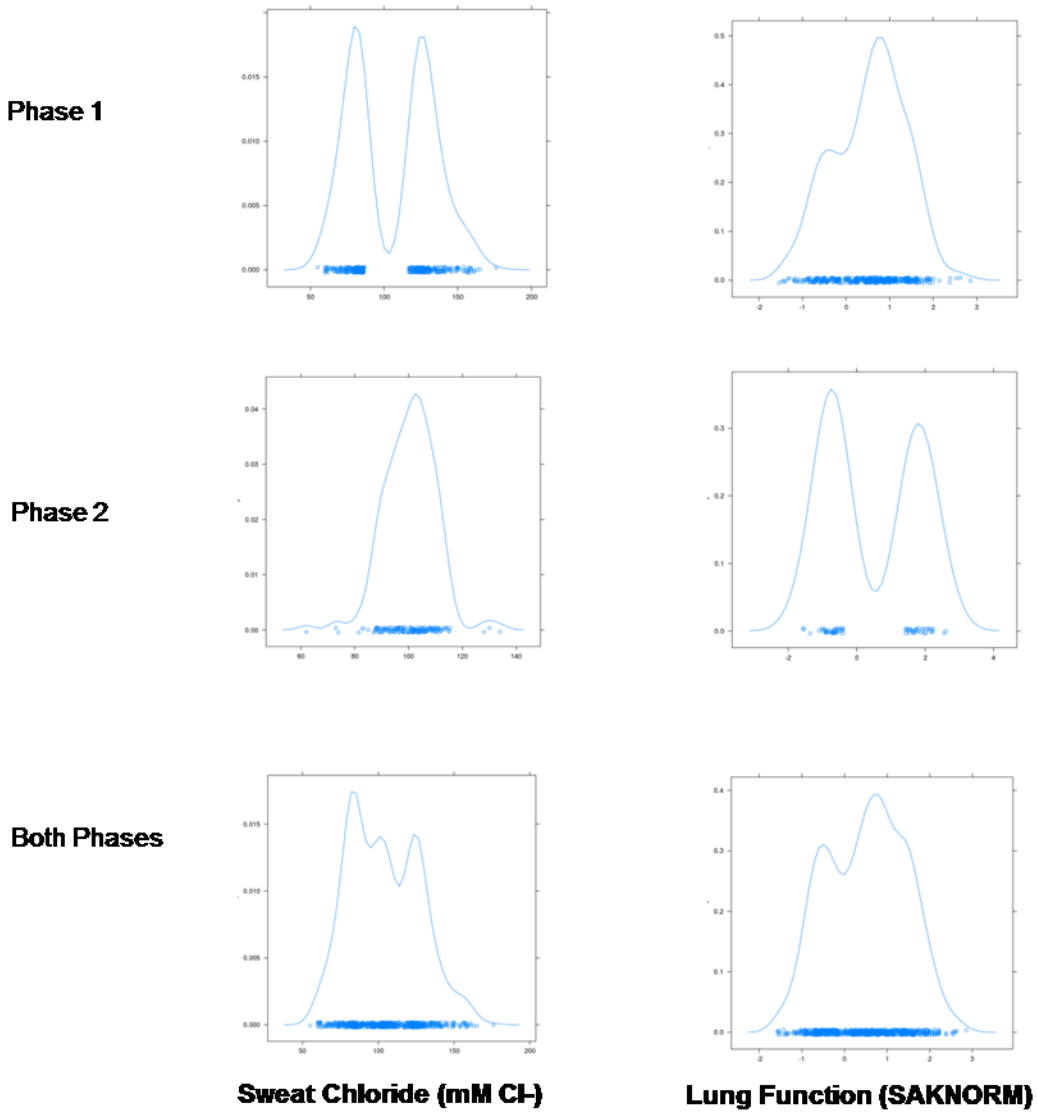


Figure 3.2. Distribution of sweat chloride concentration and lung function phenotypes by study phase. Density plots are shown to indicate distribution of trait values in either phase 1 (n=486), phase 2 (n=276), or both phases combined (n=762).

Re-sequencing Summary Statistics and Variant Annotation	
F508del # samples	762
Start/End of Phase 1 Capture	chr7:117107678-117317732
Start/End of Phase 2 Capture	chr7:116916359-117426330
Total Targeted Bases	191,621 (n=486, Original Phase 1 Capture Design)
	400,150 (n=276, Expanded Phase 2 Capture Design)
% Bases Depth > 20x	98.2% (n=486, Original Phase 1 Capture Design)
	97.6% (n=276, Expanded Phase 2 Capture Design)
Mean Depth of Coverage	176x (n=486, Original Phase 1 Capture Design)
	94x (n=276, Expanded Phase 2 Capture Design)
Ts/Tv Ratio	2.3 (n=762, Combined Captures)
Total Variants	1,388
Common (MAF > 1%)	602
Rare (MAF < 1%)	786
Private (1 chromosome)	539
Extragenic	234
CFTR	chr7:117120017-117308718
Exonic	18
Nonysynonymous	12
Synonymous	6
+/- 20 bp of canonical splice	9
Intronic	628
5' / 3' UTR	6
WNT2	chr7:116916686-116963343
ASZ1	chr7:117003276-117067577
CTTNBP2	chr7:117350706-117513561
Exonic, 5'/3' UTR	37
Intronic	465

Summary statistics are provided for the 510kb region surrounding the CFTR locus in F508del homozygous patients only. The project was completed in two phases, with separate, but semi-overlapping re-sequencing captures (see original vs expanded capture). All variant call statistics are provided for the combined capture. Intragenic variants have been divided into CFTR (blue highlight) and others (yellow highlight). All genomic coordinates are from build hg19.

Table 3.1. Re-sequencing Summary Statistics and Variant Annotation.

Exonic CFTR Variants in 762 F508del Homozygotes							
Chr7 bp (hg19)	Variant Type	Transcript Annotation	# Chromosomes	Frequency	Association Beta Value (mM Sweat Cl-)*	Association Beta Value (SAKNORM)*	rsID
117171069	synonymous	p.L130L	1	0.001			
117175331	synonymous	p.I203I	1	0.001			rs1800081
117199524	nonsynonymous	p.L467F	4	0.003	7.76	-0.002	rs1800089
117199533	nonsynonymous	p.V470M	1524	1			rs213950
117235055	synonymous	p.T854T	5	0.003	2.48	-0.31	rs1042077
117250664	nonsynonymous	p.I1027T	42	0.028	2.16	0.34	rs1800112
117304766	nonsynonymous	p.Q1330E	1	0.001			rs37566578
117305579	frameshift deletion	c.4203_4206del:p.1401_1402del	1	0.001			
117305584	frameshift insertion	c.4208_4209insCTGC:p.R1403fs	1	0.001			
117305588	synonymous	p.I1404I	1	0.001			
117307031	nonsynonymous	p.R1438Y	1	0.001			
117307108	synonymous	p.Q1463Q	35	0.023	-1.45	-0.07	rs1800136
117307142	nonsynonymous	p.V1475M	1	0.001			

*Association beta values provided for variants with > 1 chromosomes only.

Table 3.2. Exonic *CFTR* Variants in 762 F508del Homozygotes.

Mis-mapped CFTR Exonic Variation Due to Extragenic Regions of High Homology to Exon 9					
Chr7 bp (hg19)	Variant Type	Transcript Annotation	# Chromosomes	Frequency	rsID
117188715	frameshift deletion	c.1231_1235del:p.411_412del	2	0.001	
117188736	nonsynonymous	p.N417K	5	0.003	rs4727853
117188750	nonsynonymous	p.S422F	3	0.002	rs201880593
117188797	nonsynonymous	p.T438A	4	0.003	rs201434579
117188840	inframe deletion	p.L453del	1	0.001	rs377319489
117188850	synonymous	p.A455A	4	0.003	rs79074685

Table 3.3. Mis-mapped *CFTR* Exonic Variation Due to Extragenic Regions of High Homology to Exon 9.

	Chr7 Position (hg19)	Ref	Alt	Variant Type	Location	Frequency	# Ref Chrom	# Alt Chrom	Combined Capture Beta	Capture 1 Point-wise Permutation P-value	Capture 2 Point-wise Permutation P-value	Combined Point-wise Permutation P-value	Combined Max(T) Permutation P-value
Phase 2 Coverage Only	116929105	CTT	C	INDEL	Intron 4 WNT2	0.013	381	5	10.57	-	0.0174	-	0.992
	116943793	A	T	SNP	Intron 3 WNT2	0.016	380	6	-8.64	-	0.0309	-	1.000
	116950636	C	CAAA	INDEL	Intron 3 WNT2	0.088	353	31	4.74	-	0.0061	-	0.876
	117002471	A	AT	INDEL	800bp 3' of ASZ1	0.060	364	22	5.33	-	0.0128	-	0.990
	117019187	18 TA	7 TA	INDEL	Intron 10 ASZ1	0.008	383	3	-11.21	-	0.0466	-	1.000
	117044471	CT	C	INDEL	Intron 4 ASZ1	0.049	368	18	-4.96	-	0.0341	-	1.000
	117047463	TA	T	INDEL	Intron 4 ASZ1	0.940	199	187	8.26	-	0.0383	-	1.000
	117065313	C	CTTT	INDEL	Intron 2 ASZ1	0.108	31	287	4.01	-	0.0033	-	0.630
	117085677	CT	C	INDEL	35kb 5' CFTR	0.253	78	308	2.84	-	0.0396	-	1.000
	117106383	G	A	SNP	14kb 5' CFTR	0.011	380	4	-11.64	-	0.0201	-	0.995
Phase 1 & Phase 2 Overlapping Coverage	117144829	G	GAAA	INDEL	Intron 2 CFTR	0.019	375	7	-7.59	-	0.0371	0.0350	1.000
	117153491	G	GAA	INDEL	Intron 3 CFTR	0.072	1258	90	7.02	0.0091	0.1504	0.0096	0.977
	117250479	CA	C	INDEL	Intron 18 CFTR	0.640	822	526	-5.13	0.0214	0.6682	0.0202	1.000
	117252108	C	T	SNP	Intron 20 CFTR	0.022	894	20	12.83	0.0421	-	0.0421	1.000
	117258532	TCA	T	INDEL	Intron 21 CFTR	0.403	818	330	5.51	0.0042	0.1870	0.0029	0.673
	117284114	C	CT	INDEL	Intron 23 CFTR	0.016	380	6	-8.95	-	0.0256	0.0200	1.000
	117302037	23 G	24 G	INDEL	Intron 24 CFTR	0.344	1003	345	5.40	0.0047	0.9382	0.0030	0.984
	117302037	23 G	22 G	INDEL	Intron 24 CFTR	0.108	1217	131	-5.96	0.0131	0.5779	0.0106	0.699
Phase 2 Coverage Only	117357046	T	C	SNP	Intron 22 CTTNBP2	0.038	372	14	-5.31	-	0.0431	-	1.000
	117403609	CTT	C	INDEL	Intron 9 CTTNBP2	0.221	308	68	3.74	-	0.0025	-	0.597
	117419783	CAT	C	INDEL	Intron 7 CTTNBP2	0.122	344	42	3.63	-	0.0272	-	1.000

Table 3.4. Common Variant Sweat Chloride Level Associations.

Variants above 1% minor allele frequency in F508del homozygotes associating with sweat chloride levels via linear regression prior to multiple test correction (uncorrected p-value < .05, 21/602 total common variants). Overall beta values correspond to the either the combined or phase 2 only associations and indicate increase or decrease in mM Cl- units. Multiple test corrected combined study phase p-value is indicated by Max(T) permutation. Other combined and phase-specific uncorrected p-values were empirically calculated via 1x10^6 point-wise phenotypic permutations.

	Chr7 Position (hg19)	Ref	Alt	Variant Type	Location	Frequency	# Ref Chrom	# Alt Chrom	Combined Capture Beta	Capture 1 Point-wise Permutation P-value	Capture 2 Point-wise Permutation P-value	Combined Point-wise Permutation P-value	Combined Max(T) Permutation P-value
Phase 2 Coverage Only	116933160	T	TA	INDEL	Intron 4 WNT2	0.010	483	5	1.12	-	0.0067	-	0.924
	116941856	CTT	C	INDEL	Intron 3 WNT2	0.224	410	92	0.24	-	0.0270	-	1.000
	116944486	ATTT	A	INDEL	Intron 3 WNT2	0.056	462	26	-0.41	-	0.0332	-	1.000
	117012943	A	G	SNP	Intron 10 ASZ1	0.017	484	8	-0.69	-	0.0083	-	0.981
	117033209	T	TA	INDEL	Intron 4 ASZ1	0.056	462	26	-0.51	-	0.0076	-	0.949
	117081552	21T	23T	INDEL	14.5kb 5' ASZ1	0.038	470	18	0.52	-	0.0209	-	1.000
	117081552	21T	19T	INDEL	14kb 5' ASZ1	0.054	463	25	-0.41	-	0.0347	-	1.000
	117106419	C	CTT	INDEL	13.5kb 5' CFTR	0.084	450	38	0.33	-	0.0408	-	1.000
Phase 1 & Phase 2 Overlapping Coverage	117138870	G	GTTTT	INDEL	Intron 1 CFTR	0.069	829	57	-0.28	0.0175	-	0.0176	0.999
	117144829	G	A	SNP	Intron 2 CFTR	0.111	1233	137	0.17	0.4222	0.0207	0.0366	1.000
	117152687	CAA	C	INDEL	Intron 3 CFTR	0.233	1114	260	-0.20	0.0781	0.0402	0.0041	0.810
	117158623	CAT	C	INDEL	Intron 3 CFTR	0.294	1045	307	-0.13	0.1869	0.0458	0.0435	1.000
	117160319	17T	18T	INDEL	Intron 3 CFTR	0.472	931	439	0.20	0.0510	0.1169	0.0037	0.771
	117160319	17T	16T	INDEL	Intron 3 CFTR	0.197	1145	225	-0.21	0.0889	0.0602	0.0034	0.738
	117198607	C	CTAAATAAA	INDEL	Intron 10 CFTR	0.074	94	1270	0.21	0.0162	0.4595	0.0207	1.000
	117206653	TA	T	INDEL	Intron 11 CFTR	0.021	1346	28	-0.37	0.6681	0.0137	0.0287	1.000
	117226828	CT	C	INDEL	Intron 11 CFTR	0.078	1274	100	0.20	0.2810	0.0219	0.0313	1.000
	117250664*	T	C (I1027T)	SNP	Exon 19 CFTR	0.032	1332	42	0.34	0.0364	0.2103	0.0130	0.995
	117252077	G	GTATA	INDEL	Intron 20 CFTR	0.225	976	220	0.13	0.1791	0.1412	0.0360	1.000
	117278790	C	CATAA	INDEL	Intron 22 CFTR	0.024	1293	31	-0.29	0.2221	0.0855	0.0435	1.000
Phase 2 Coverage Only	117363806	CTTT	C	INDEL	Intron 19 CTTNBP2	0.038	470	18	0.46	-	0.0416	-	1.000

Table 3.5. Common Variant Lung Function Associations.

Variants above 1% minor allele frequency in F508del homozygotes associating with a change in lung function, as assessed by SAKNORM (22/602 total common variants). Positive beta values indicate increased lung function. * Indicates I1027T coding variant.

Abbreviated	Full Reference
Smith and Dekker 2016	Smith, E.M., Lajoie, B.R., Jain, G., and Dekker, J. (2016). Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus. Am J Hum Genet 98, 185-201.
Yang and Harris 2015	Yang, R., Kerschner, J.L., Gosalia, N., Neems, D., Gorsic, L.K., Safi, A., Crawford, G.E., Kosak, S.T., Leir, S.H., and Harris, A. (2015). Differential contribution of cis-regulatory elements to higher order chromatin structure and expression of the CFTR locus. Nucleic Acids Res.

Chromosome	Start (hg19)	End (hg19)	References	Additional Annotation
chr7	117035669	117046178	Smith and Dekker 2016 / Yang and Harris 2015	-80kb
chr7	117046179	117050387	Smith and Dekker 2016	
chr7	117050388	117057168	Smith and Dekker 2016	
chr7	117057169	117062301	Smith and Dekker 2016	
chr7	117075245	117076245	Smith and Dekker 2016 / Yang and Harris 2015	-44kb
chr7	117078458	117079183	Smith and Dekker 2016	
chr7	117079184	117082828	Smith and Dekker 2016	
chr7	117082829	117083304	Smith and Dekker 2016	
chr7	117084910	117086709	Yang and Harris 2015	-35kb
chr7	117094392	117100244	Smith and Dekker 2016 / Yang and Harris 2015	-20.9kb
chr7	117120896	117125750	Smith and Dekker 2016	
chr7	117129829	117130579	Yang and Harris 2015	exon1+10kb
chr7	117133646	117135039	Smith and Dekker 2016	
chr7	117139373	117140740	Smith and Dekker 2016	
chr7	117167209	117172399	Smith and Dekker 2016	
chr7	117172400	117179549	Smith and Dekker 2016	
chr7	117211555	117213631	Smith and Dekker 2016 / Yang and Harris 2015	intron11ab
chr7	117220582	117223082	Yang and Harris 2015	intron 11c
chr7	117224217	117229090	Smith and Dekker 2016 / Yang and Harris 2015	intron 12
chr7	117237162	117245849	Smith and Dekker 2016	
chr7	117255414	117258789	Smith and Dekker 2016	
chr7	117273186	117274601	Smith and Dekker 2016	
chr7	117296108	117301295	Smith and Dekker 2016	
chr7	117302414	117304652	Smith and Dekker 2016	
chr7	117304653	117304780	Smith and Dekker 2016	
chr7	117305161	117306850	Smith and Dekker 2016	
chr7	117307580	117308782	Smith and Dekker 2016	
chr7	117326535	117332210	Smith and Dekker 2016	
chr7	117332211	117336769	Smith and Dekker 2016	
chr7	117336770	117337958	Smith and Dekker 2016	
chr7	117337959	117341506	Smith and Dekker 2016	
chr7	117355221	117357250	Yang and Harris 2015	+48.9kb

Table 3.6. Regions Found to Interact with CFTR Promotor via Chromatin Capture by either Dekker or Harris groups.

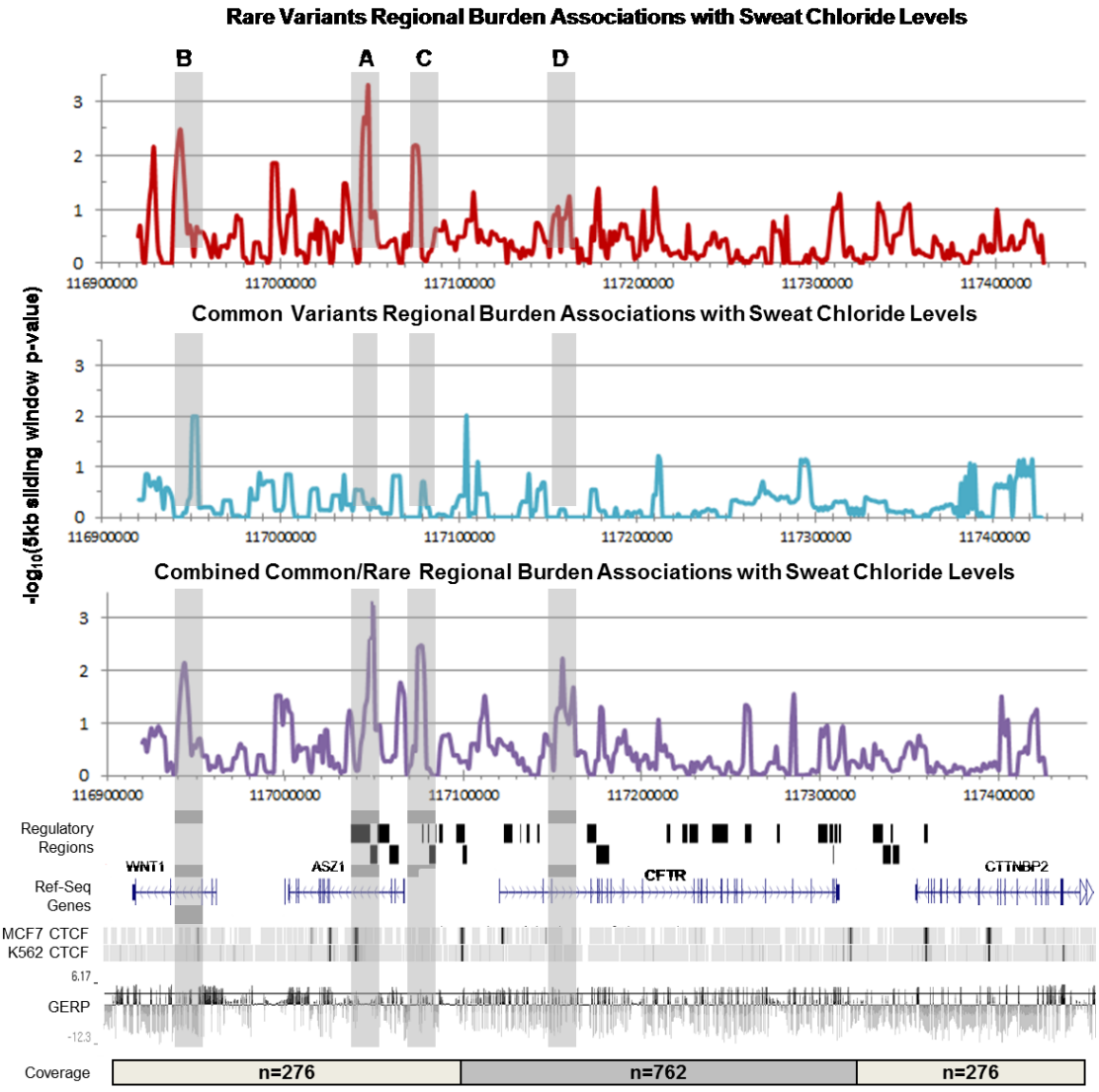


Figure 3.3. Burden Testing of Common and Rare Variants Associating with Sweat Cl- Levels.

All variants within each 5kb window, moved across the entire region in increments of 1250bp, were tested for a combined association with sweat chloride levels (mM) via SKAT-O test. X-axis denotes chr7 hg19 bp position, y-axis is –log10 of the regional p-value. Association values were plotted at the center of each 5kb window. Top (Red): Rare variants with MAF < 1% only. Middle (Blue): Common variants with MAF > 1%

only. Bottom (Purple): Combined test of common and rare variants with variants weighted inversely proportional to their frequency. Vertical shaded boxes: regions of significant association in the combined analysis (α=.01). Genome browser style tracks: Top, packed view of known *CFTR* regulatory regions of interest as previously reported (see supplemental file). Middle, view of genes with exonic/intronic structure. Bottom, CTCF binding signals in two cell types, and mammalian conservation as assayed by GERP (horizontal bar indicating GERP score of 4).

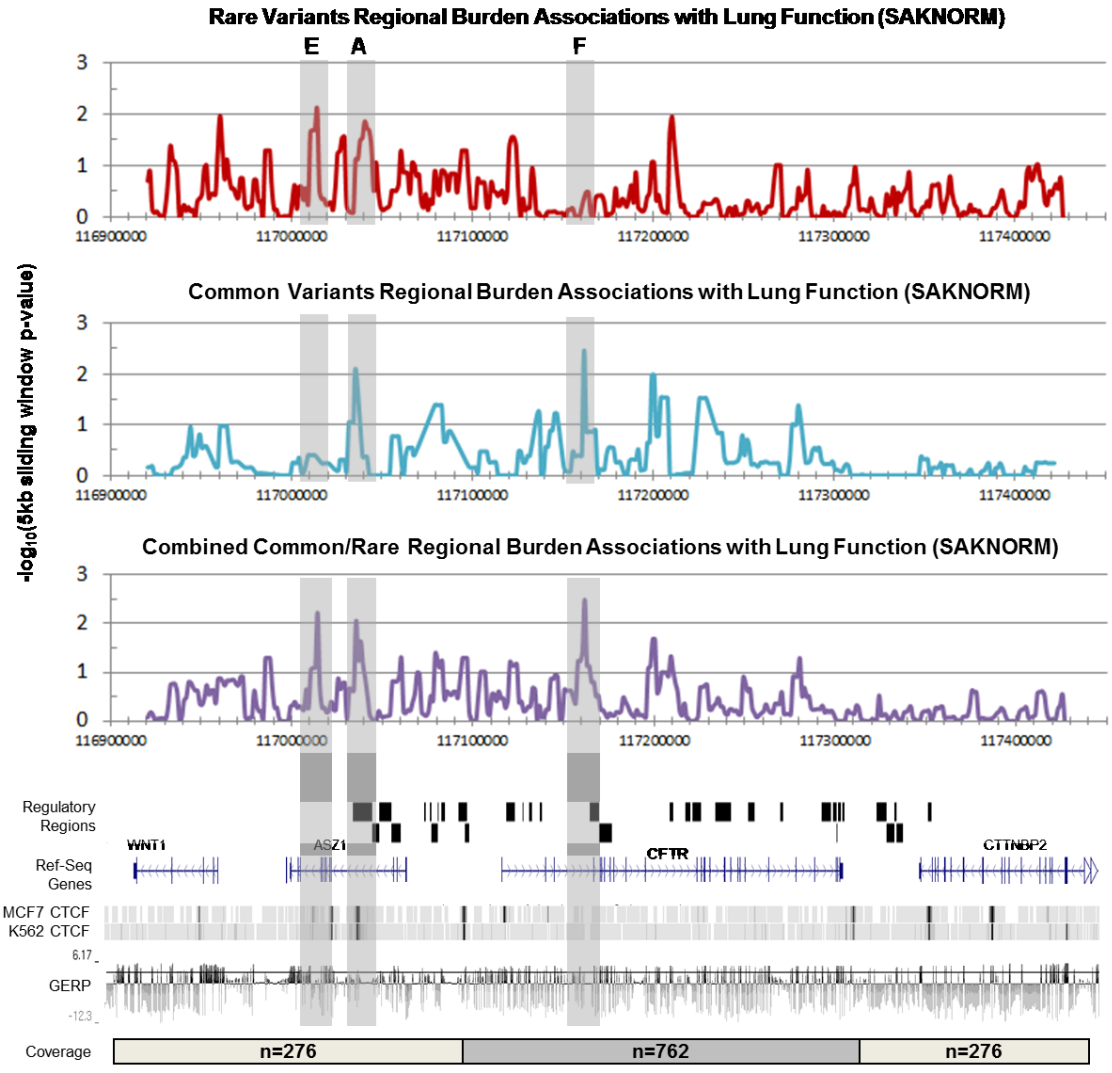


Figure 3.4. Burden Testing of Common and Rare Variants Associating with Lung Function (SAKNORM).

All variants within each 5kb window, moving across the entire region in increments of 1250bp, were tested for a combined association with lung function (SAKNORM) via SKAT-O test. X-axis denotes chr7 hg19 bp position, y-axis is –log10 of the regional p-value. Association values were plotted at the center of each 5kb window. Top (Red): Rare variants with MAF < 1% only. Middle (Blue): Common variants with MAF > 1%

only. Bottom (Purple): Combined test of common and rare variants with variants weighted inversely proportional to their frequency. Vertical shaded boxes: regions of significant association in the combined analysis (α=.01).Genome browser style tracks: Top, packed view of known *CFTR* regions of interest as previously reported (see supplemental file). Middle, view of genes with exonic/intronic structure. Bottom: CTCF binding signals in two cell types, and mammalian conservation as assayed by GERP (horizontal bar indicating GERP score of 4).

Region A					
Variant	**# Chr Allele 1**	**# Chr Allele 2**	**Frequency**	**Variant Beta (Mm Cl-)**	**Uncorrected Variant P-value**
7:117044382:C:CT	194	96	0.495	-2.522	0.06313
7:117044382:CT:C	258	64	0.248	1.068	0.4603
7:117044471:C:CTT	366	10	0.027	2.128	0.4888
7:117044471:CT:C	350	18	0.051	-4.962	0.03325
7:117045434:C:T	384	1	0.003	9.349	0.3242
7:117046808:C:CA	264	61	0.231	1.231	0.4008
7:117046808:CA:C	268	59	0.220	-1.027	0.4875
7:117047463:TA:T	12	187	0.064	8.264	0.03432
7:117047463:TAA:T	382	2	0.005	23.24	0.0004516
7:117048043:G:A	384	1	0.003	-12.26	0.1956
7:117048087:C:T	384	1	0.003	2.313	0.8076
7:117050389:T:TTA	384	1	0.003	2.313	0.8076
7:117050798:GA:G	382	2	0.005	16.47	0.01365
7:117052093:A:AGTGTG	354	16	0.045	0.5097	0.8367
7:117052093:AGT:A	384	1	0.003	-19.3	0.041
7:117052094:G:GTGTGT	2	192	0.010	4.724	0.6188
7:117052096:G:GTGTA	14	186	0.075	0.4438	0.9031
7:117052098:G:GTA	326	30	0.092	1.195	0.5252
7:117052100:G:A	370	6	0.016	2.153	0.585
7:117052102:G:A	384	1	0.003	-19.3	0.041
7:117052846:A:C	384	1	0.003	4.223	0.6565
Region B					
Variant	**# Chr Allele 1**	**# Chr Allele 2**	**Frequency**	**Variant Beta (Mm Cl-)**	**Uncorrected Variant P-value**
7:116941856:C:CT	338	24	0.071005917	-3.087	0.134
7:116941856:C:CT	380	3	0.007894737	-17.3	0.001467
7:116941856:C:CT	382	2	0.005235602	-3.233	0.631
7:116941856:C:CTT	340	27	0.079411765	0.6333	0.7087
7:116941856:C:CTTT	350	18	0.051428571	0.01316	0.9955
7:116941856:C:CTTTT	348	19	0.054597701	0.8883	0.6978
7:116941856:CT:C	242	72	0.297520661	-0.9596	0.4958
7:116941856:CTT:C	258	68	0.263565891	1.33	0.3105
7:116942115:G:A	380	3	0.007894737	-17.3	0.001467
7:116942433:G:T	380	3	0.007894737	-17.3	0.001467
7:116942912:G:A	372	7	0.018817204	-0.9478	0.7949
7:116943281:C:A	382	2	0.005235602	2.527	0.7073
7:116943793:A:T	374	6	0.016042781	-8.634	0.02696
7:116944283:T:A	380	3	0.007894737	-17.3	0.001467
7:116944481:T:C	378	4	0.010582011	5.26	0.271
7:116944484:A:C	360	12	0.033333333	-0.3475	0.9023
7:116944486:AT:A	162	62	0.382716049	0.01824	0.9922
7:116944486:ATT:A	148	119	0.804054054	-1.513	0.2797
7:116944486:ATTT:A	354	16	0.04519774	-0.2978	0.9041
7:116944486:ATTTTTTT	360	13	0.036111111	0.4875	0.8578
7:116945583:G:T	384	1	0.002604167	12.36	0.1919
7:116947196:A:C	366	10	0.027322404	-3.743	0.2227
7:116947789:C:A	384	1	0.002604167	-13.27	0.1613
7:116947844:A:T	362	12	0.033149171	-2.034	0.471
7:116948171:C:A	372	7	0.018817204	-5.617	0.1223
Region C					
Variant	**# Chr Allele 1**	**# Chr Allele 2**	**Frequency**	**Variant Beta (Mm Cl-)**	**Uncorrected Variant P-value**
7:117074330:A:G	384	1	0.003	11.36	0.2307
7:117074834:A:T	384	1	0.003	-8.242	0.385
7:117076029:G:A	382	2	0.005	-20.92	0.001648
7:117076523:G:A	384	1	0.003	-4.724	0.6188
7:117077757:A:T	384	1	0.003	13.37	0.1581
7:117078936:T:A	384	1	0.003	-11.76	0.2146
Region D					
Variant	**# Chr Allele 1**	**# Chr Allele 2**	**Frequency**	**Variant Beta (Mm Cl-)**	**Uncorrected Variant P-value**
7:117153275:A:G	960	1	0.001	-30.3	0.2741
7:117153491:G:GAA	1168	90	0.077	7.023	0.009601
7:117153491:GA:G	1330	9	0.007	-2.329	0.7725
7:117153761:A:AT	1338	5	0.004	27.87	0.009528
7:117154574:G:A	960	1	0.001	16.8	0.5446
7:117155052:G:GT	894	34	0.038	2.813	0.5681
7:117155422:A:G	944	3	0.003	13.48	0.3992
7:117155926:T:C	960	1	0.001	-30.3	0.2741
7:117156287:AT:A	380	3	0.008	-0.4055	0.9413

Table 3.7. Variants within regions associating with sweat chloride levels via SKAT.

Region A					
Variant	**# Chr Allele 1**	**# Chr Allele 2**	**Frequency**	**Variant Beta (SAKNORM)**	**Uncorrected Variant P-value**
7:117032902:C:CT	312	88	0.282	-0.000694	0.9955
7:117032902:CT:C	206	141	0.684	-0.02552	0.8314
7:117032902:CTT:C	468	10	0.021	-0.1418	0.6349
7:117033209:TA:T	246	121	0.492	0.1024	0.387
7:117033209:T:TA	436	26	0.060	-0.5075	0.007731
7:117036759:A:G	482	3	0.006	-1.097	0.04039
7:117041435:GT:G	220	134	0.609	0.01611	0.8923
7:117041435:GTT:G	406	41	0.101	0.1244	0.432
7:117041448:T:A	480	4	0.008	0.8106	0.08123
7:117043459:C:T	480	4	0.008	0.1836	0.6938
7:117044382:C:CT	236	126	0.534	-0.01676	0.8876
7:117044382:CT:C	326	81	0.248	0.4209	0.07924
7:117044471:C:CTT	462	13	0.028	-0.02575	0.9222
7:117044471:CT:C	440	24	0.055	0.008147	0.9673
7:117046808:CA:C	350	69	0.197	-0.05393	0.6817
7:117046808:C:CA	332	78	0.235	0.04221	0.7396
Region B					
Variant	**# Chr Allele 1**	**# Chr Allele 2**	**Frequency**	**Variant Beta (SAKNORM)**	**Uncorrected Variant P-value**
7:117010858:TTG:T	458	15	0.033	-0.4275	0.08197
7:117010858:TTGTG:T	482	3	0.006	-0.1513	0.7782
7:117010858:T:TTG	26	238	0.109	-0.05116	0.8053
7:117010858:T:TTGTG	426	31	0.073	-0.02861	0.8722
7:117010858:T:TTGTGT(	474	7	0.015	-0.3241	0.3605
7:117011094:G:A	468	10	0.021	0.1029	0.7304
7:117012943:A:G	476	8	0.017	-0.6888	0.01022
Region C					
Variant	**# Chr Allele 1**	**# Chr Allele 2**	**Frequency**	**Variant Beta (SAKNORM)**	**Uncorrected Variant P-value**
7:117160308:G:T	848	19	0.022	0.1404	0.4742
7:117160319:G:GT	931	439	0.472	0.2003	0.003651
7:117160319:GT:G	1145	225	0.197	-0.2067	0.003311
7:117160319:GTT:G	470	7	0.015	0.216	0.5411
7:117163435:G:GT	100	393	0.254	0.1275	0.31
7:117163435:G:GTT	240	323	0.743	-0.07606	0.395
7:117164446:GA:G	1210	82	0.068	-0.1514	0.1385
7:117164446:G:GA	194	590	0.329	0.09749	0.3057

Table 3.8. Variants within regions associating with lung function via SKAT.

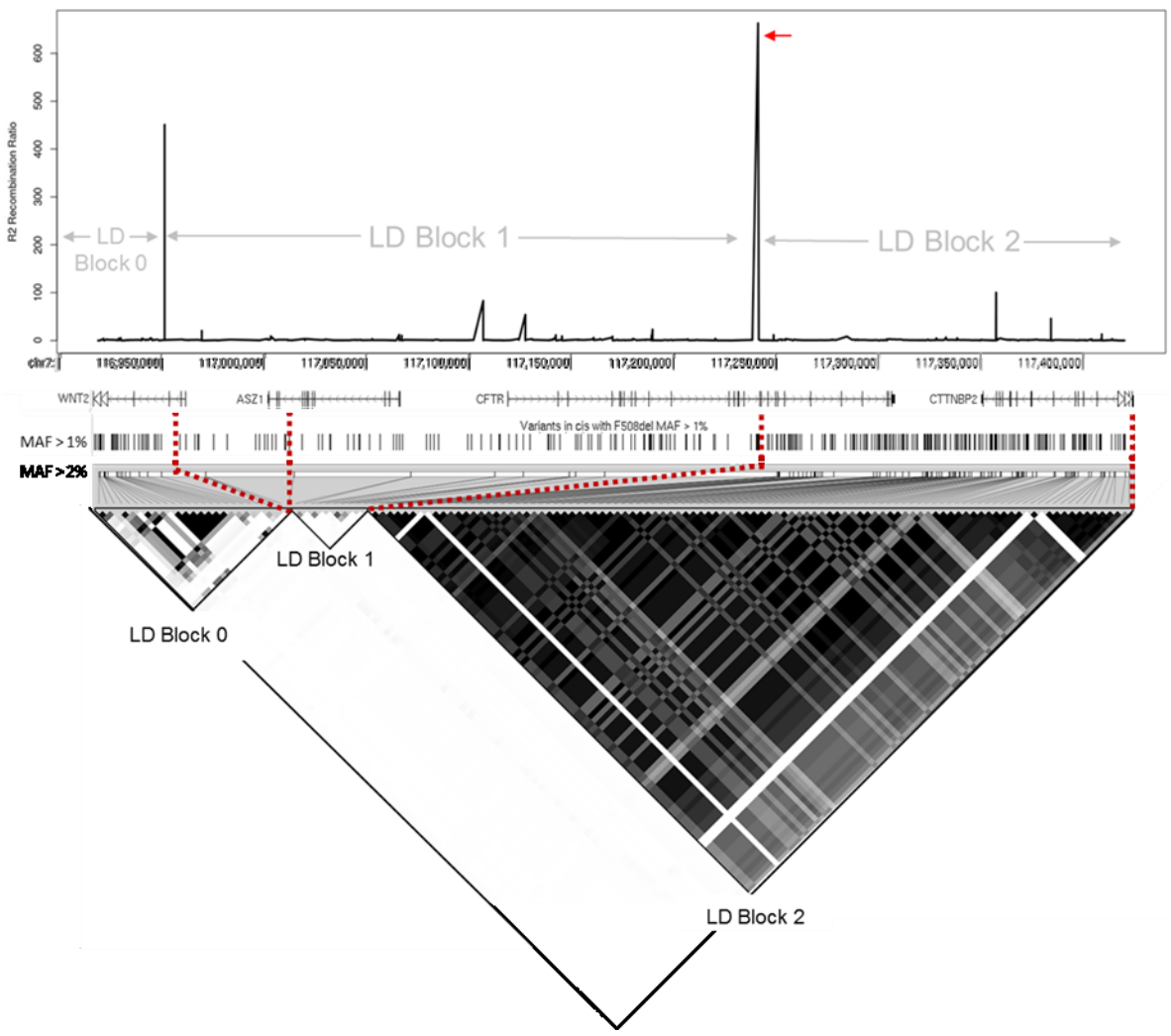


Figure 3.5. Recombination ratio and linkage disequilibrium observed in 762 F508del homozygous samples (1524 chromosomes) across a 506 kb re-sequencing region surrounding *CFTR*.

Top– Recombination ratio plotted by genomic location. A recombination event occurring within intron 15 of *CFTR* indicated by red arrow. Below is an intronic and exonic map of known RefSeq genes, and re-sequencing study variants with MAF > 1% (hg19 coordinates). Bottom- LD heat map of variants with MAF > 2% in the F508del population. Dashed red lines indicate projection of variants from their genomic positions

to the heat-map of r^2 values below. Within the heatmap, black boxes indicate an r^2 value of 1 or complete LD, while white boxes indicate an r^2 of 0 or linkage equilibrium. Three proposed LD blocks are outlined (triangles). The first extends from the start of the sequencing capture to intron 3 of WNT2. LD block 1 then extends from the WNT2 locus to intron 15 of *CFTR*. And finally, LD block 2 extends from intron 15 of *CFTR* to the end of the sequencing capture (mid-CTTNBP2). LD blocks 0 and 2 likely extend far beyond the capture design

.

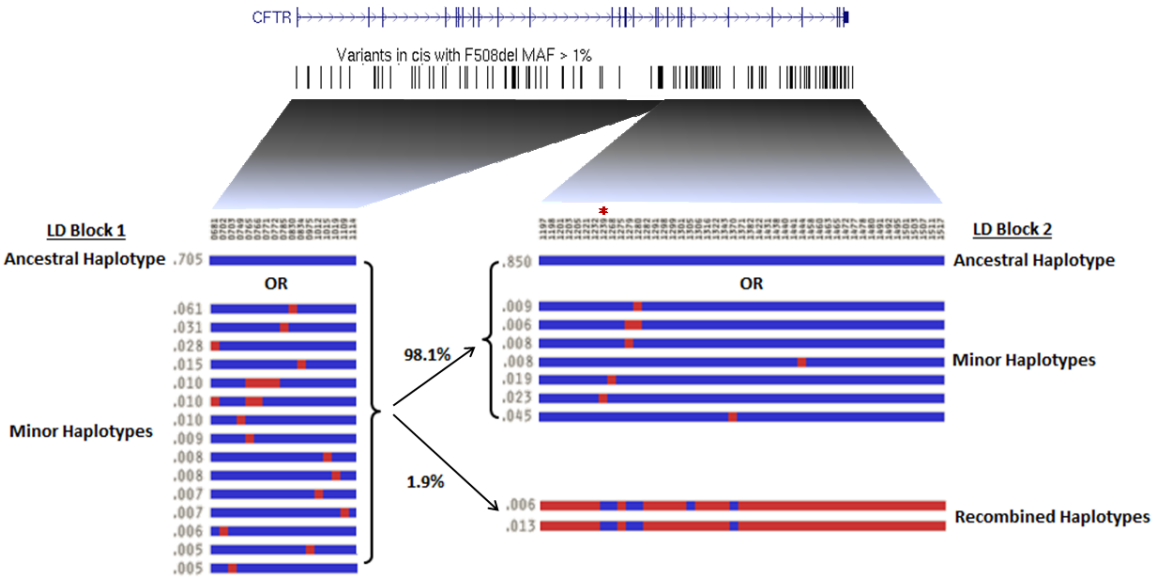


Figure 3.6.

Haplotypes observed in 762 F508del homozygous samples (1524 chromosomes) across the CFTR locus.

Top– Intronic and exonic map of *CFTR*, and contained re-sequencing study variants with MAF > 1%. Bottom – Representation of *CFTR* SNP haplotypes with MAF > 1% and MHF > 0.5%. Each numbered variant is represented vertically by blue (reference allele) or red (alternate allele) squares. Each haplotype is represented horizontally as a row, with respective minor haplotype frequencies to the left. Two LD blocks are shown, with primary recombination events indicated by bold connecting lines. LD block 1 shows one ancestral haplotype with minor haplotypes below, while LD block 2 shows an additional recombined haplotype.

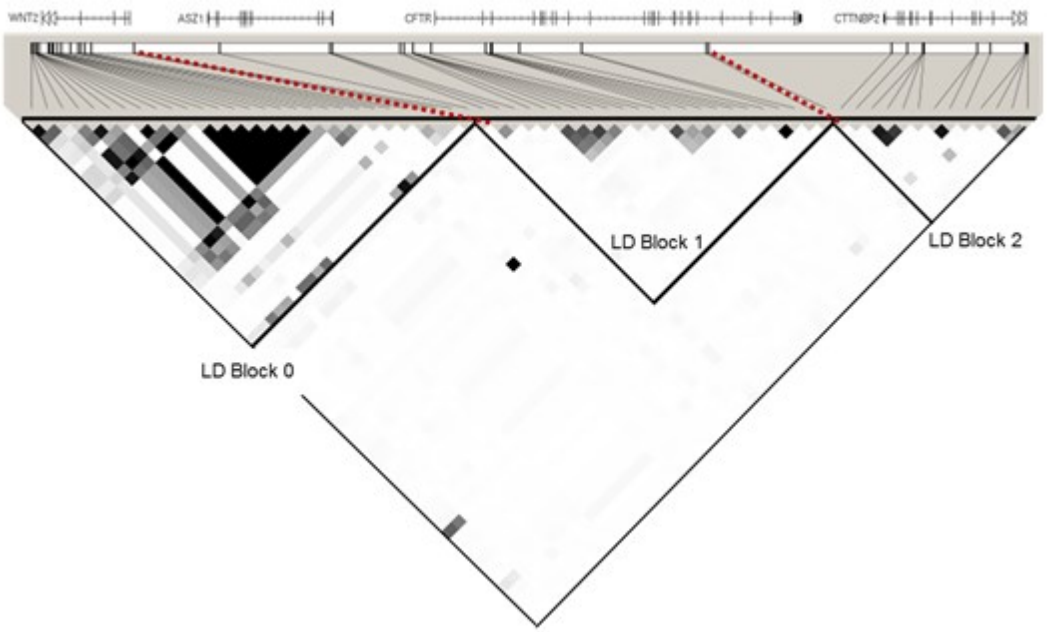


Figure 3.7

Linkage disequilibrium across a 506 kb re-sequencing region surrounding CFTR after removing samples with alternative ancestral haplotype in LD block 2.

Top– Intronic and exonic map of known RefSeq genes, and re-sequencing study variants with MAF > 1% (hg19 coordinates). Bottom- LD heat map of variants with MAF > 2% in F508del chromosomes lacking the intron 15 recombination event. Dashed red lines indicate projection of variants from their genomic positions to the heat-map below. Within the heatmap, black boxes indicate an r^2 value of 1 or complete LD, while white boxes indicate an r^2 of 0 or linkage equilibrium. The three proposed LD blocks are outlined (triangles). The first extends from the start of the sequencing capture to intron 3 of WNT2. LD blocks one and two are now indistinguishable, indicating the alternative ancestral haplotype in LD block 2 delineates the recombination event within intron 15.

LD Block 1

31 SNPs capturing 99.2% of all SNPs with MAF > 1% with r2 correlation value > 0.9

Gray highlight: 13 SNPs capturing all haplotypes with a minor haplotype frequency (MHF) > 1%

		Marker #	671	681	702	703	749	765	766	771	772	783	785	830	834	974	975	983	1012	1015	1019	1109	1114
MHF		**SNP**	7:117115127	7:117119881	7:117126947	7:117126950	7:117138870	7:117141826	7:117141830	7:117141840	7:117141847	7:117144828	7:117144832	7:117158642	7:117160308	7:117189107	7:117189137	7:117189272	7:117194040	7:117194626	7:117196675	7:117219746	7:117220436
0.575			C	T	C	G	G	G	C	C	C	G	C	G	G	T	T	A	G	T	G	T	G
0.051			C	T	C	G	G	G	C	C	C	G	C	A	G	T	T	A	G	T	G	T	G
0.036			C	T	C	G	G	G	C	C	C	A	C	G	G	T	T	A	G	T	G	T	G
0.032			C	T	C	G	G	G	C	C	C	G	C	G	G	C	T	A	G	T	G	T	G
0.025			C	T	C	G	G	G	C	C	C	A	A	G	G	T	T	A	G	T	G	T	G
0.020			C	G	C	G	G	G	C	C	C	G	C	G	G	T	T	A	G	T	G	T	G
0.013			C	T	C	G	G	G	C	C	C	G	C	G	T	T	T	A	G	T	G	T	G
0.013			C	T	C	G	G	G	C	C	C	G	C	G	G	T	T	G	G	T	G	T	G
0.013			C	T	C	G	G	G	C	C	C	G	C	G	G	T	T	G	G	T	G	A	G
0.011			C	T	C	G	G	A	A	A	A	G	C	G	G	T	T	A	G	T	G	T	G
0.010			C	G	C	G	G	A	A	C	C	G	C	G	G	T	T	A	G	T	G	T	G
0.008			C	T	C	G	G	A	C	C	C	G	C	G	G	T	T	A	G	T	G	T	G
0.008			T	T	C	G	G	G	C	C	C	G	C	G	G	T	T	A	G	T	G	T	G
0.008			C	T	C	G	G	G	C	C	C	G	C	G	G	T	T	A	G	T	A	T	G
0.006			C	T	C	G	G	G	C	C	C	G	C	G	G	T	T	A	G	C	G	T	G
0.006			C	T	C	G	G	G	C	C	C	G	C	G	G	C	T	G	G	T	G	T	G
0.006			C	T	C	G	G	G	C	C	C	G	C	G	G	C	T	G	G	T	G	A	G
0.006			C	T	C	G	T	G	C	C	C	G	C	G	G	T	T	A	G	T	G	T	G
0.005			C	T	C	G	G	G	C	C	C	G	C	G	G	T	T	A	A	T	G	T	G
0.005			C	T	C	A	G	G	C	C	C	G	C	G	G	T	T	A	G	T	G	T	G
0.005			C	T	C	G	G	G	C	C	C	G	C	G	G	C	T	A	G	T	G	A	G
0.004			C	T	C	G	G	G	C	C	C	G	C	G	G	C	C	A	G	T	G	T	G
0.004			C	T	G	A	G	G	C	C	C	G	C	G	G	T	T	A	G	T	G	T	G
0.003			C	T	C	G	G	G	C	C	C	G	C	G	G	T	T	A	G	T	G	T	T
0.003			C	T	C	G	G	G	C	C	C	G	C	A	G	T	T	G	G	T	G	T	G
0.003			C	T	C	G	G	G	C	C	C	G	C	A	G	T	T	G	G	T	G	A	G
0.003			C	T	C	G	G	G	C	C	C	G	C	G	G	T	T	A	G	T	G	A	G
0.003			C	T	C	G	G	A	A	C	C	G	C	G	G	T	T	A	G	T	G	T	G
0.003			C	T	C	G	G	G	C	C	C	A	C	G	G	T	T	G	G	T	G	T	G
0.003			C	T	C	G	G	G	C	C	C	A	C	G	G	T	T	G	G	T	G	A	G
0.003			C	T	C	G	G	A	A	C	C	G	C	A	G	T	T	A	G	T	G	T	G
0.003			C	T	C	G	G	G	C	A	C	G	C	G	G	T	T	A	G	T	A	T	G
0.003			C	T	G	G	G	G	C	C	C	G	C	G	G	T	T	A	G	T	G	T	G
0.002			C	T	C	G	G	G	C	C	C	G	C	A	G	C	C	A	G	T	G	T	G
0.002			C	T	C	G	T	G	C	C	C	G	C	A	G	T	T	A	G	T	G	T	G
0.002			C	T	C	G	G	A	A	A	C	G	C	G	G	T	T	A	G	T	A	T	G
0.002			T	G	C	G	G	G	C	C	C	G	C	G	G	T	T	A	G	T	G	T	G
0.002			C	G	C	G	G	G	C	C	C	A	A	G	G	T	T	A	G	T	G	T	G
0.002			C	T	C	G	T	G	C	C	C	A	C	G	G	T	T	A	G	T	G	T	G
0.002			C	T	C	G	G	A	A	A	C	G	C	A	G	T	C	A	G	T	G	T	G
0.002			C	T	C	G	G	G	C	C	C	A	C	G	G	C	T	A	G	T	G	T	G
0.002			C	T	C	G	G	A	C	C	C	G	C	A	G	T	T	A	G	T	G	T	G
0.002			C	T	C	G	G	G	C	C	C	G	C	A	G	T	T	A	G	T	G	A	G
0.002			C	T	C	G	G	G	C	C	C	A	C	G	G	T	T	A	G	C	G	T	G
0.002			C	T	C	G	G	G	C	C	C	G	C	A	G	C	T	G	G	T	G	T	G
0.002			C	T	C	G	G	G	C	C	C	G	C	A	G	C	T	G	G	T	G	A	G
0.001			C	T	C	G	G	G	C	C	C	G	A	G	G	T	T	A	G	T	G	T	G
0.001			T	T	C	G	G	G	C	C	C	G	C	A	G	T	T	A	G	T	G	T	G
0.001			C	G	C	G	G	G	C	C	C	G	C	G	G	C	T	A	G	T	G	T	G
0.001			C	T	C	G	T	G	C	C	C	G	C	G	G	T	T	A	A	T	G	T	G
0.001			C	T	C	G	G	G	C	C	C	G	C	A	G	C	T	A	G	T	G	T	G
0.001			C	T	C	G	G	G	C	C	C	G	C	A	G	T	T	A	G	T	G	A	T
0.001			C	T	C	G	G	G	C	C	C	A	A	G	T	C	T	A	G	T	G	T	G
0.001			C	T	C	G	G	G	C	C	C	A	C	G	G	C	T	A	G	T	G	A	G
0.001			C	T	C	G	G	G	A	A	C	G	C	G	G	T	T	A	G	T	A	T	G
0.001			C	T	C	G	G	G	C	C	C	G	C	A	G	T	T	A	A	T	G	T	G
0.001			C	T	C	G	G	G	C	C	C	A	C	G	G	C	T	G	G	T	G	T	G
0.001			C	T	C	G	G	G	C	C	C	A	C	G	G	C	T	G	G	T	G	A	G

LD Block 2

		Marker #	1239	1268	1271	1279	1280	1301	1305	1370	1444	1475	1477
MHF		**SNP**	7:117250664	7:117252077	7:117252108	7:117254085	7:117254093	7:117259011	7:117259340	7:117273998	7:117289573	7:117299434	7:117300188:A:C
0.828			T	G	C	A	C	G	A	T	A	A	A
0.045			T	G	C	A	C	G	A	G	A	A	A
0.025			C	G	C	A	C	G	A	T	A	A	A
0.020			T	G	T	A	C	G	A	T	A	A	A
0.019			T	A	C	A	C	G	A	T	A	A	A
0.009			T	G	C	A	T	G	A	T	A	A	A
0.008			T	G	C	G	C	G	A	T	A	A	A
0.006			T	G	C	A	C	G	A	T	A	G	A
0.005			T	G	C	A	C	G	A	T	T	A	A
0.004			T	G	C	G	T	G	A	T	A	A	A
0.003			T	G	C	A	C	A	G	T	T	G	C
0.003			T	G	T	A	C	A	G	T	T	G	C
0.003			T	G	C	A	C	A	A	T	A	G	A
0.003			C	G	C	A	C	G	A	G	A	A	A
0.003			T	G	C	A	C	A	A	T	T	G	C
0.003			T	G	T	A	C	A	A	T	T	G	C
0.002			T	A	C	A	C	G	A	G	A	A	A
0.002			C	G	C	G	C	G	A	T	A	A	A
0.001			T	G	T	A	C	G	A	T	T	A	A
0.001			T	G	C	G	T	G	A	T	A	G	A
0.001			T	A	C	A	C	G	A	T	T	A	A
0.001			T	G	C	A	C	G	A	T	T	G	A
0.001			T	G	T	A	C	G	A	T	A	G	A

Table 3.9. ***CFTR*** **Common Variant Haplotypes.**

Tagging SNP	Alleles Captured
7:117115127:C:T	7:117115127:C:T
7:117119881:T:G	7:117119881:T:G
7:117126947:C:G	7:117126947:C:G
7:117126950:G:A	7:117126950:G:A
7:117138870:G:T	7:117138870:G:T
7:117141826:G:A	7:117141826:G:A
7:117141830:C:A	7:117141830:C:A
7:117141840:A:C	7:117141840:A:C
7:117141847:C:A	7:117141847:C:A
7:117144828:G:A	7:117144828:G:A
7:117144832:C:A	7:117144832:C:A
7:117158642:A:G	7:117158642:A:G
7:117160308:G:T	7:117160308:G:T
7:117189107:T:C	7:117189107:T:C,7:117189105:T:C
7:117189137:T:C	7:117189137:T:C
7:117189272:G:A	7:117189272:G:A
7:117194040:G:A	7:117194040:G:A
7:117194626:T:C	7:117194626:T:C
7:117196675:G:A	7:117196675:G:A
7:117219746:T:A	7:117219746:T:A
7:117220436:G:T	7:117220436:G:T
7:117300188:A:C	7:117241859:T:C,7:117288078:G:A,7:117305374:G:A,7:117313645:C:T,7:117300259:C:G,7:117259946:C:T,7:117312350:T:G,7:117307108:G:A,7:117248234:T:G,7:117300188:A:C,7:117256712:G:A,7:117303145:C:A,7:117311713:T:G,7:117293552:T:G,7:117311123:A:G,7:117309974:G:A,7:117267511:C:A,7:117277926:G:A,7:117246636:G:A,7:117292696:G:A,7:117254527:A:C,7:117241508:A:G,7:117311066:A:G,7:117262304:G:A,7:117294301:C:T,7:117300742:G:A,7:117252874:G:A,7:117311117:C:
7:117250664:T:C	7:117250664:T:C
7:117252077:G:A	7:117252077:G:A
7:117252108:C:T	7:117252108:C:T
7:117254085:A:G	7:117254085:A:G
7:117254093:C:T	7:117254093:C:T
7:117259340:A:G	7:117259340:A:G
7:117273998:T:G	7:117273998:T:G
7:117259011:G:A	7:117289539:G:T,7:117241031:C:A,7:117311978:T:C,7:117258529:G:A,7:117261293:C:T,7:117258372:A:G,7:117259011:G:A,7:117286524:G:A,7:117240668:A:T,7:117289540:T:A,7:117289332:A:G
7:117289573:A:T	7:117289573:A:T
7:117299434:G:A	7:117291132:A:G,7:117277554:A:T,7:117299434:G:A,7:117246078:G:A

Table3.10. Tagging of *CFTR* SNPs with MAF > 1%, r^2 > 0.9.

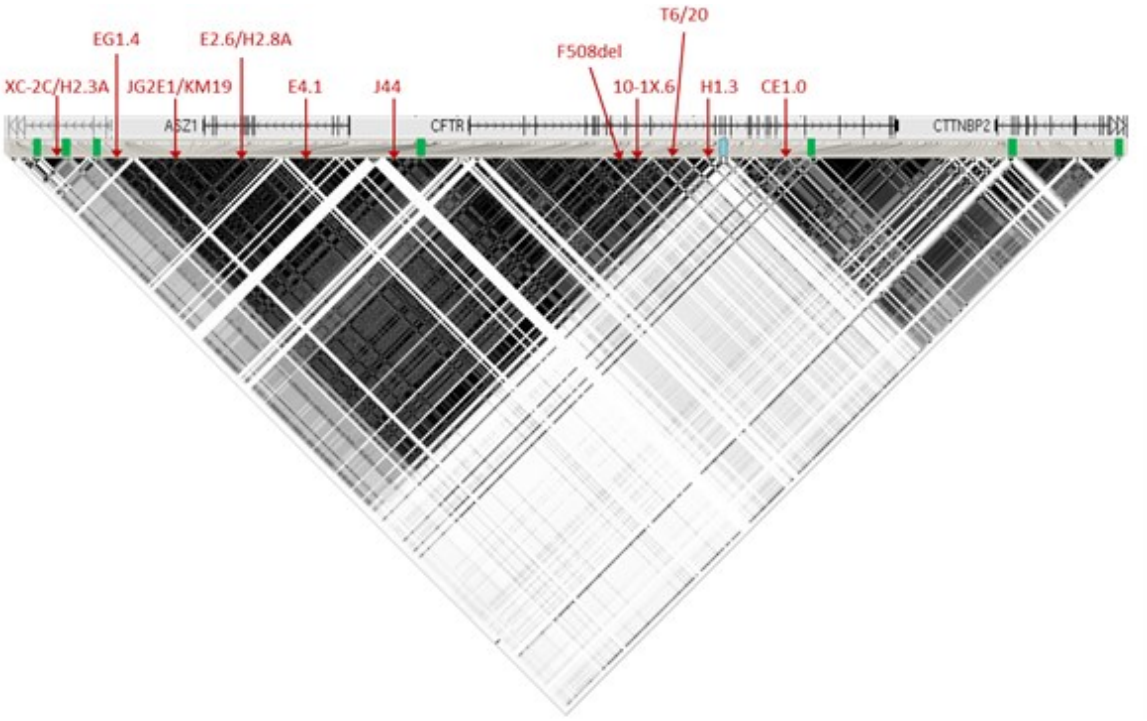


Figure 3.8. Linkage disequilibrium observed in 762 F508del homozygous samples and 163 alternative allele samples (1850 chromosomes total) across a 506kb region surrounding *CFTR*.

Top– Red arrows indicate approximate locations of RFLPs originally used to map *CFTR* Middle- Intronic and exonic map of known RefSeq genes. Bottom- LD heat map of variants with MAF > 5% in the combined population of F508del homozygotes and alternative genotype samples. Within the heatmap, black boxes indicate an r^2 value of 1 or complete LD, while white boxes indicate an r^2 of 0 or linkage equilibrium. One primary recombination event in intron 15 is visualized (blue box), and additional recombination events of lower frequency are observed (green boxes).

Chapter 4

Significance and Future Directions

Carbonic Anhydrase XII Deficiency

In **Chapter 2** of this thesis, I described the discovery of disease causing variants within *CA12* in three patients with hyponatremic dehydration and elevated sweat chloride levels. While this gene had previously been implicated in a similar phenotype in a consanguineous population, our study describes three novel loss-of-function variants, two of which were inherited in a compound heterozygous fashion in a non-consanguineous patient. This research has advanced our understanding of both the *CA12* deficiency phenotype and the molecular etiology of this disease state. Finally, it has provided key support for the hypothesis that bicarbonate imbalance may play a substantial role in CF lung disease.

The Mild Lung Phenotype in *CA12* Deficiency

A novel finding in **Chapter 2** is the observation of lung disease in a patient with loss of function mutations in *CA12*. We understand there may be unrelated or additional causes of lung disease in this singleton, and address these possibilities here. Patients with CA12 deficiency were likely only directed to our study due to their distinct elevated sweat chloride levels, which suggest a CF diagnosis. While one patient reported in **Chapter 2** had mild bronchiectasis and some lung infections consistent with mild CF, it is possible these symptoms would have been overlooked by a primary care physician had her sweat test been within a normal range. Indeed, everyone is susceptible to viral and bacterial infections which are generally treated at home and never formally cultured. Perhaps this individual would have been given more time to clear infections had she been viewed as a healthy individual, as opposed to a CF patient. We note that the mild lung

disease in this patient was not observed in any other *CA12* deficient patient. This could be for various reasons including: reduced vigilance for these symptoms in the other pedigrees (they were not seen regularly at a CF clinic), a younger age (perhaps lung disease manifests itself later), or altered presentation due to different causative mutations (allele specific expressivity – to be addressed below).

One reason for the potential variable expressivity of lung disease in *CA12* patients may be the inheritance of alternative disease causing variants. A key finding reported in **Chapter 2** was that both the previously reported mutation and the mutations presented in our study led to complete loss of CA XII function. Previously, it was reported that the initial mutation only resulted in a 30% reduction in enzymatic activity (*86*). However, we believe that this original assay was erroneous in that it failed to consider the salt concentrations at the active site of the protein (which is extracellular). After correcting for the salt concentration, both mutations led to loss of function. This highlights the importance of a thorough and thoughtful investigation into the molecular etiology of any given candidate disease causing variant.

We additionally show in our study the correct trafficking and localization of CA XII protein resulting from the reported missense mutations in a polarized cell line (MDCK). During revisions and publication of our manuscript, another group reported that one of these mutations (E143K), resulted in altered glycosylation of the protein and retention in the ER in MDCK cells (*187*). While our initial findings suggested possible altered glycosylation states of the E134K mutant, this appeared to have no effect on protein localization. These discrepant findings could be due to an altered expression of proteins in these pathways between polarized and non-polarized cells. Alternatively, the

vectors expressing these supposedly identical *CA12* transcripts may somehow play a role here. An ideal way to resolve this question would be to obtain primary cells from a patient harboring this mutation, which may not be possible. Sharing vectors and repeating these assays would be a more feasible and warranted endeavor, such that hopefully one finding is replicable. While it is interesting to posit that these two protein expressing mutations may have alternative paths to loss of function, the phenotypes in these two pedigrees are largely concordant in that both lack mild lung disease trait. However, the proband with lung disease happens to carry mutations which result in little to no protein expression. Perhaps the residual protein in the other pedigrees somehow staves off lung disease, but this seems highly unlikely as our functional studies have shown lack of enzymatic activity in these proteins even when they reach the cell surface. I posit that the more likely reason for the variable expressivity of the lung phenotype in *CA12* deficiency is either a delayed onset or incomplete phenotyping.

As noted in **Chapter 2,** the mechanism by which loss of CA XII function results in elevated sweat chloride levels remains to be fully elucidated. The recent paper by Hong and Muhammad also found that CA XII may activate the chloride/bicarbonate exchanger AE2 as part of a larger metabolon in the sweat duct (*187*). Additional studies will be needed to determine other potential members of this ion modulating complex in the sweat duct, and their role in maintenance of sweat chloride concentration.

Additional *CA12* Carrier Cohorts for Future Studies

Future studies into *CA12* deficiency may benefit from recruiting patients with similar phenotypes to that reported in **Chapter 2**. If lung disease does develop in these

patients, it is possible that there are additional *CA12* deficient patients which could be ascertained through the study of idiopathic bronchiectasis patients. Our laboratory is currently planning to recruit and screen potential *CA12* variant carriers from this large population (~110,000 patients in the US) (*114*). The other key presentations which led to ascertainment of *CA12* deficient patients were failure to thrive (FTT) and hyponatremic dehydration. For the oldest proband in our study, failure to thrive in infancy led to sweat testing and a CF diagnosis. While sweat testing is indicated in patients with failure to thrive, it may not be ordered by all clinicians or practices. Thus, it is possible that some children with FTT may carry disease causing variants within *CA12,* and these patients have not been ascertained to CF clinics or our study. The other proband presented in **Chapter 2** was initially sweat tested due to an episode of hyponatremic dehydration. There are 1.5 million pediatric outpatient visits per year in the U.S. due to dehydration. Of these, approximately 3% can be further classified as hyponatremic dehydration (*188*). Given that sweat testing is not necessarily indicated for dehydration, it is possible that additional *CA12* deficient patients are present in this population. Of note, the pedigree reported in **Chapter 2** lives in an arid climate, where one might expect an increased incidence of dehydration in the general population. Heatwaves and strenuous activity can often lead to general dehydration, but some individuals may be more or less susceptible to developing these symptoms. Interestingly, a heat-wave in 1948 led to an increased frequency of admission of CF patients to pediatric wards for symptoms of dehydration (*3*). This increase in admissions was likely due to the increased salt wasting in CF coupled with the high surface area to volume ratio of small children (*189*). Given that salt wasting is also a predominant feature of *CA12* deficiency, it seems highly plausible

there are additional affected individuals who may present with dehydration only during certain extremities of weather or exercise. The latter may be justified by a recent study of hyponatremia in triatheletes (*190*). Here, the authors showed that ~10% of these athletes developed hyponatremia, and another 0.3% developed critical hyponatremia. There athletes may be carriers of *CA12* disease causing variation, or may have similar genetic alterations in genes of related pathways (including *CFTR*). This is yet another source where additional insights into the pathways related to salt wasting could be gained.

A Novel Role for Bicarbonate in CF

This study has challenged the notion that cystic fibrosis is solely a disease of chloride transport. Our work reveals that a CF-like phenotype in both the sweat gland and lung can arise from aberrant bicarbonate levels. The renown CF researcher and patient, Dr. Paul Quinton, has hypothesized that altered bicarbonate transport by CFTR may be driving pathogenesis in CF (*99*). In 2008, he suggested that in the CF lung, excess bicarbonate (due to its role as a buffer) sequesters cations required for the proper expansion of mucins. This theory has since been tested and proven correct in both mouse and pig animal models of CF (*191, 192*). Additionally, others have shown that the lung ASL pH in both CF patients and other animal models is significantly reduced (*193*). This change in pH has been linked to an altered host defense and an increase in infections (*194*). Future studies into the role of bicarbonate in CF pathogenesis will require a thorough understanding of bicarbonate regulatory pathways, and how these vary among tissues. The work presented in **Chapter 2** has added to and will continue to inform this growing body of knowledge.

The Atypical Cystic Fibrosis Study

There are additional patients with elevated sweat chloride levels in the absence of *CFTR* mutations who have been recruited to the atypical CF sequencing study. Like our *CA12* samples, some of these probands and families have had whole exome or whole genome sequencing analysis in hopes of elucidating causative variation. Unfortunately, these assays have not yielded many candidate genes for the elevated sweat chloride phenotype. Clinical exome sequencing has a diagnostic rate between 10 and 30% (*195*). The variability here arises from many sources including the degree of previous genetic analysis, the distinct phenotype, the genetic power (larger pedigrees are more informative), and the NGS diagnostic criteria.

The low yield of candidate genes in the atypical CF study may be a combination of three possible sources. First, a large number of candidate variants and genes do not meet the high threshold of genetic and bioinformatic evidence needed to move forward into functional studies. A candidate gene must have two likely loss of function mutations that are inherited in a manner consistent with the pedigree's phenotype, be expressed in the tissues of interest (primarily the sweat gland and lung), and play a role in relevant pathways (i.e. ion transport, secretion, etc). Many candidate genes in the atypical study meet only two out of these three criteria. It is possible that mutations in multiple genes which may be interacting in a pathway could lead to disease in these patients (poly or multi-genic inheritance) (*196*). However, the methods to assay this disease mechanism are limited and still not well understood (*197*). Re-analysis of this cohort in a pathway specific manner may yield additional candidates as the tools become available, and pathway curation improves.

Like other genome sequencing studies, NGS analysis in the atypical CF study is contingent upon a reliable phenotype and distinct inheritance pattern. We have observed multiple families where sweat chloride levels have changed over time in the same patient, or a pattern of inheritance within the family is unclear. Often, individuals may have borderline sweat chloride levels and could easily be labelled as either affected or unaffected. Variability in sweat chloride levels is a known phenomenon, but has not been well characterized. Some factors which may influence sweat chloride levels include hormone cycles (*198*), salt intake and hydration, and variability that is inherent with the sweat testing method. There is an ongoing effort to understand this variability in the Cutting laboratory, and future publications will hopefully elucidate some of these mechanisms. Additionally, there is currently an active collaboration to develop a wearable sweat chloride sensor that would record these intra-individual changes in real time. This additional research will help further refine and increase the overall usability of the sweat chloride phenotype.

Finally, the atypical CF sequencing study may simply not be assaying the causative variation in these patients. Exomes only assay the ~1% of the genome which we believe to be functional. While there is significant evidence this is a good filter when looking for disease causing variation, there are still additional types of variation that can lead to disease. These include deep intronic splice variants, copy number variation, and modification of regulatory regions. Additional sources of heritable factors include epigenetic factors and mitochondrial DNA. As the cost of whole genome sequencing decreases, there will be further opportunities to assay extra-exonic regions in this cohort.

However, bioinformatics methods to query these datasets are still in active development, and it will likely be many years before we can realize their full potential.

Towards Understanding Phenotypic Variability in CF

Chapter 3 described a targeted resequencing study of the *CFTR* locus in a specific cohort of CF patients. This work makes several contributions to the understanding of genetic modifiers in CF. We reported novel variation *in cis* with the F508del allele, as well as variation within nearby blocks of linkage disequilibrium. These findings challenged the previous notion that the F508del homozygous population was highly homogeneous in regards to the *CFTR* locus. We also describe a recombination event, refined in this study to intron 15 of *CFTR*, which contributes a significant portion of the common variation observed. This recombination event appears to be specific to the F508del population, suggesting it arose after the founder mutation. Finally, and perhaps most interestingly, some of the variation in this cohort was found to associate with variability in two key CF traits: sweat chloride levels and lung function. While other rare revertants have been described (see **Chapter 1**), this is the first report of *CFTR* variation beyond the F508del mutation modifying either lung function or sweat chloride levels in a large cohort.

Nearly all bioinformatics assays of NGS biomedical research data are hypothesis generating work, and this study is no exception. The rare and common variation identified in these studies will be further assayed for mechanisms by which they may alter *CFTR* expression of function. First, some of the rare coding variation reported *in*

cis with F508del could be assayed *in vitro* for revertant potential. We have already shared some of these variants with collaborators working on *CFTR* transcriptional rate studies, in order to determine whether the alternative codon usage could ultimately lead to protein misfolding. Next, many additional studies will be needed to further dissect the associations observed during region based burden testing. In many cases, the primary hypothesis to be resolved at these loci will be whether the rare and common variation could alter CTCF binding or overall chromatin structure. There has been extensive work defining the *CFTR* topologically associated domain in recent years by three key groups (*148, 156, 168*). It is my hope that this study has now provided these groups with an additional toolset with which to assay mechanisms regulating expression of the *CFTR* protein. As mentioned in **Chapter 3,** the findings presented should be tested for replication in a larger cohort. Often, studying small to moderate sized cohorts can lead to a higher frequency of false positive findings through population stratification and other methods. Many means of replication exist, which could include genotyping common variants under a given burden test peak, or alternatively genotyping all variation under each peak through a similar targeted sequencing method. Ideally, these assays would be conducted in a novel cohort, potentially recruited at alternative sites.

While this work now introduces the idea of genetic modifiers within and near the *CFTR* locus, there have been ongoing efforts to elucidate genome-wide variation associated with CF lung function through GWAS (see **Chapter 1**). Beyond the scope of this thesis, there were additional previously identified modifier loci for both lung disease and CF related diabetes sequenced in the cohort presented in **Chapter 3**. Each locus will require a similar analysis in order to elucidate potential causative variation that may be

present under these large GWAS peaks. However, many of these loci were only typed in phase 2 of the study, so the cohort size is quite small. Future work in the Cutting lab in collaboration with other sites includes whole genome sequencing of a large CF patient cohort (~5000). This study will allow for not only replication of the study in presented in **Chapter 3**, but will allow for a genome-wide assay of rare variation which may contribute to the phenotypic heterogeneity in this population. By recognizing the mechanisms which lead to mild disease in select patients, these studies may inform future CF therapies.

References

1. C. F. Foundation, "Cystic Fibrosis Foundation Patient Registry annual Data Report 2012," (Bethesda, MD, 2013).
2. G. Mehta, M. Macek, A. Mehta, Cystic fibrosis across Europe: EuroCareCF analysis of demographic data from 35 countries. *J Cyst. Fibros* **9**, S5-S64 (2010).
3. J. M. Littlewood, D. G. Peckham, Ed. (Cystic Fibrosis Medicine).
4. J. Gonzalo-Ruiz *et al.*, Early determination of cystic fibrosis by electrochemical chloride quantification in sweat. *Biosensors and Bioelectronics* **24**, 1788-1791 (2009).
5. P. Lebecque *et al.*, Mutations of the cystic fibrosis gene and intermediate sweat chloride levels in children. *Am J Respir. Crit Care Med* **165**, 757-761 (2002).
6. S. Gee, C. A. Herter, W. K. Dicke, On the celiac disease. *St Bart Hosp Rep* **24**, 17-20 (1888).
7. D. H. Andersen, Cystic fibrosis of the pancreas and its relation to celiac disease: a clinical and pathologic study. *Am. J. Dis. Child* **56**, 344-399 (1938).
8. J. Katkin *et al.*, Cystic fibrosis: Assessment and management of pancreatic insufficiency. *UpToDate. April* **12**, (2013).
9. D. Crozier, Cystic fibrosis: a not-so-fatal disease. *Pediatric Clinics of North America* **21**, 935 (1974).
10. J. R. Crossley, R. B. Elliott, P. A. Smith, Dried-blood spot screening for cystic fibrosis in the newborn. *Lancet* **1**, 472-474 (1979).
11. D. Holsclaw, H. Eckstein, H. Nixon, Meconium ileus: a 20-year review of 109 cases. *American Journal of Diseases of Children* **109**, 101-113 (1965).
12. D. H. Anderson, Cystic fibrosis of the pancreas and its relation to celiac disease. *Am J Dis Child* **56**, 344-399 (1938).
13. P. E. di SANT'AGNESE, D. H. Andersen, Celiac syndrome: IV. Chemotherapy in infections of the respiratory tract associated with cystic fibrosis of the pancreas; observations with penicillin and drugs of the sulfonamide group, with special reference to penicillin aerosol. *American Journal of Diseases of Children* **72**, 17-61 (1946).
14. H. Shwachman, L. L. Kulczycki, Long-term study of one hundred five patients with Cystic Fibrosis. *Am. J. Dis. Child* **96**, 6-15 (1958).
15. A. L. Stephenson *et al.*, A contemporary survival analysis of individuals with cystic fibrosis: a cohort study. *European Respiratory Journal* **45**, 670-679 (2015).
16. J. R. WILSON, R. DuBois, Report of a fatal case of keratomalacia in an infant, with postmortem examination. *American Journal of Diseases of Children* **26**, 431-446 (1923).
17. K. D. Blackfan, S. B. Wolbach, Vitamin A deficiency in infants: a clinical and pathological study. *The Journal of Pediatrics* **3**, 679-706 (1933).
18. K. W.R., D. H. Andersen, **Heat prostration in fibrocystic disease of the pancreas and other condition.** *Pediatrics* **8**, **648** (1951).
19. R. C. Darling, A. Paul, G. A. Perera, D. H. Andersen, ELECTROLYTE ABNORMALITIES OF THE SWEAT IN FIBROCYSTIC DISEASE OF THE PANCREAS [degrees]. *The American journal of the medical sciences* **225**, 67-70 (1953).
20. L. E. Gibson, R. E. Cooke, A test for concentration of electrolytes in sweat in cystic fibrosis of the pancreas utilizing pilocarpine by iontophoresis. *Pediatrics* **23**, 545-549 (1959).
21. P. M. Quinton, Physiological basis of cystic fibrosis: a historical perspective. *Physiol Rev* **79**, S3-S22 (1999).
22. U. Hopfer, P. Will, D. Dearborn, in *Perspectives in Cystic Fibrosis: Proceedings of the Eighth Internation Congress on CF Toronto*. (1980).

23. M. R. Knowles, J. Gatzy, R. Boucher, Increased bioelectric potential difference across respiratory epithelia in cystic fibrosis. *N. Engl. J. Med* **305**, 1489-1495 (1981).
24. P. M. Quinton, J. Bijman, Higher bioelectric potentials due to decreased chloride absorption in the sweat glands of patients with cystic fibrosis. *N. Engl. J. Med* **308**, 1185-1189 (1983).
25. R. A. Frizzell, G. Rechkemmer, R. L. Shoemaker, Altered regulation of airway epithelial cell chloride channels in cystic fibrosis. *Science* **233**, 558-560 (1986).
26. J. R. Riordan, CFTR Function and Prospects for Therapy. *Annu. Rev. Biochem*, 701-726 (2008).
27. L. C. Tsui *et al.*, Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* **239**, 1054-1057 (1985).
28. R. White *et al.*, A closely linked genetic marker for cystic fibrosis. *Nature* **318**, 382-384 (1985).
29. B. J. Wainwright *et al.*, Localization of cystic fibrosis locus to human chromosome 7cen-q22. *Nature* **318**, 384-385 (1985).
30. R. G. Knowlton *et al.*, A polymorphic DNA marker linked to cystic fibrosis is located on chromosome 7. *Nature* **318**, 380-382 (1985).
31. G. R. Cutting *et al.*, Analysis of DNA polymorphism haplotypes linked to the cystic fibrosis locus in North American Black and Caucasian families supports the existence of multiple mutations of the cystic fibrosis gene. *Am. J. Med. Genet* **44**, 307-318 (1989).
32. B. Kerem *et al.*, Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073-1080 (1989).
33. J. R. Riordan *et al.*, Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066-1073 (1989).
34. J. M. Rommens *et al.*, Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**, 1059-1065 (1989).
35. P. R. Sosnay *et al.*, Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat. Genet* **45**, 1160-1167 (2013).
36. E. Kerem *et al.*, The relation between genotype and phenotype in cystic fibrosis--analysis of the most common mutation (deltaF508). *N. Engl. J. Med* **323**, 1517-1522 (1990).
37. G. R. Cutting, Cystic fibrosis genetics: from molecular understanding to clinical application. *Nat Rev Genet* **16**, 45-56 (2015).
38. C. M. Farinha, M. D. Amaral, Most F508del-CFTR is targeted to degradation at an early folding checkpoint and independently of calnexin. *Mol Cell Biol* **25**, 5242-5252 (2005).
39. . (2015).
40. H. Yu *et al.*, Ivacaftor potentiation of multiple CFTR channels with gating mutations. *J. Cyst. Fibros* **11**, 237-245 (2012).
41. G. Borgo *et al.*, Phenotypic intrafamilial heterogeneity in cystic fibrosis. (Letter). *Clin Genet* **44**, 48-49 (1993).
42. G. Santis, L. Osborne, R. A. Knight, M. E. Hodson, Independent genetic determinants of pancreatic and pulmonary status in cystic fibrosis. *Lancet* **336**, 1081-1084 (1990).
43. M. Algire *et al.*, Analysis of CF Twins reveals that increased gene sharing does not correlate with greater similarity in lung function. *Pediatric Pulmonology Supplement* **27**, 222 (2004).
44. F. Mekus *et al.*, Categories of deltaF508 homozygous cystic fibrosis twin and sibling pairs with distinct phenotypic characteristics. *Twin. Res* **3**, 277-293 (2000).
45. I. Bronsveld *et al.*, Chloride conductance and genetic background modulate the cystic fibrosis phenotype of Delta F508 homozygous twins and siblings. *J. Clin. Invest* **108**, 1705-1715 (2001).
46. L. L. Vanscoy *et al.*, Heritability of lung disease severity in cystic fibrosis. *Am J Respir. Crit Care Med* **175**, 1036-1043 (2007).

47. J. M. Collaco, S. M. Blackman, J. McGready, K. M. Naughton, G. R. Cutting, Quantification of the Relative Contribution of Environmental and Genetic Factors to Variation in Cystic Fibrosis Lung Function. *J Pediatr* **157**, 802-807 (2010).
48. S. M. Blackman *et al.*, Relative contribution of genetic and nongenetic modifiers to intestinal obstruction in cystic fibrosis. *Gastroenterology* **131**, 1030-1039 (2006).
49. J. M. Collaco *et al.*, Effect of temperature on cystic fibrosis lung disease and infections: a replicated cohort study. *PLoS. ONE* **6**, e27784 (2011).
50. M. T. Henry *et al.*, An alpha(1)-antitrypsin enhancer polymorphism is a genetic modifier of pulmonary outcome in cystic fibrosis. *Eur. J Hum. Genet* **9**, 273-278 (2001).
51. D. D. Frangolias *et al.*, Alpha 1-antitrypsin deficiency alleles in cystic fibrosis lung disease. *Am. J. Respir. Cell Mol. Biol* **29**, 390-396 (2003).
52. M. Gabolde, M. Guilloud-Bataille, J. Feingold, C. Besmond, Association of variant alleles of mannose binding lectin with severity of pulmonary disease in cystic fibrosis: cohort study. *BMJ* **319**, 1166-1167 (1999).
53. H. Grasemann *et al.*, Airway nitric oxide levels in cystic fibrosis patients are related to a polymorphism in the neuronal nitric oxide synthase gene. *Am. J. Respir. Crit Care Med* **162**, 2172-2176 (2000).
54. A. Henrion-Caude *et al.*, Liver disease in pediatric patients with cystic fibrosis is associated with glutathione S-transferase P1 polymorphism. *Hepatology* **36**, 913-917 (2002).
55. J. N. Hirschhorn, M. J. Daly, Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95-108 (2005).
56. F. A. Wright *et al.*, Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2. *Nature Genetics* **43**, 539-546 (2011).
57. H. Corvol *et al.*, Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nat Commun* **6**, 8382 (2015).
58. S. M. Blackman *et al.*, Genetic modifiers of cystic fibrosis-related diabetes. *Diabetes* **62**, 3627-3635 (2013).
59. L. Sun *et al.*, Multiple apical plasma membrane constituents are associated with susceptibility to meconium ileus in individuals with cystic fibrosis. *Nat. Genet* **44**, 562-569 (2012).
60. T. A. Manolio *et al.*, Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).
61. O. Zuk, E. Hechter, S. R. Sunyaev, E. S. Lander, The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* **109**, 1193-1198 (2012).
62. B. C. Brown, A. Price, N. Patsopoulos, N. Zaitlen, Local joint testing improves power and identifies missing heritability in association studies. *bioRxiv*, 040089 (2016).
63. X. Chen *et al.*, Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models. *The American Journal of Human Genetics* **97**, 708-714 (2015).
64. O. Zuk *et al.*, Searching for missing heritability: Designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A* **111**, E455-E464 (2014).
65. M. Beaudoin *et al.*, Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis. *PLoS Genet* **9**, e1003723 (2013).
66. S. L. Martiniano, J. E. Hoppe, S. D. Sagel, E. T. Zemanick, Advances in the diagnosis and treatment of cystic fibrosis. *Adv Pediatr* **61**, 225-243 (2014).

67. D. Trujillano *et al.*, Next generation diagnostics of cystic fibrosis and CFTR-related disorders by targeted multiplex high-coverage resequencing of CFTR. *J. Med Genet* **50**, 455-462 (2013).
68. B. J. Rosenstein, G. R. Cutting, The diagnosis of cystic fibrosis: A consensus statement. *J. Pediatr* **132**, 589-595 (1998).
69. R. C. Stern *et al.*, Intermediate-range sweat chloride concentration and Pseudomonas bronchitis. *JAMA* **239**, 2676-2680 (1978).
70. A. Augarten *et al.*, Mild cystic fibrosis and normal or borderline sweat test in patients with the 3849+10 kb C->T mutation. *Lancet* **342**, 25-26 (1993).
71. T. V. Strong *et al.*, Cystic fibrosis gene mutation in two sisters with mild disease and normal sweat electrolyte levels. *N. Engl. J. Med* **325**, 1630-1634 (1991).
72. J. D. Groman, M. E. Meyer, R. W. Wilmott, P. L. Zeitlin, G. R. Cutting, Variant cystic fibrosis phenotypes in the absence of CFTR mutations. *N. Engl. J. Med* **347**, 401-407 (2002).
73. J. D. Groman *et al.*, Phenotypic and genetic characterization of patients with features of "nonclassic" forms of cystic fibrosis. *J Pediatr* **146**, 675-680 (2005).
74. M. B. Sheridan *et al.*, Mutations in the beta subunit of the epithelial Na+ channel in patients with a cystic fibrosis-like syndrome. *Hum. Mol. Genet* **14**, 3493-3498 (2005).
75. M. M. Reddy, M. J. Light, P. M. Quinton, Activation of the epithelial Na+ channel (ENaC) requires CFTR Cl- channel function. *Nature* **402**, 301-304 (1999).
76. K. Kunzelmann, R. Schreiber, R. Nitschke, M. Mall, Control of epithelial Na+ conductance by the cystic fibrosis transmembrane conductance regulator. *Pflugers Arch* **440**, 193-201 (2000).
77. S. S. Chang *et al.*, Mutations in subunits of the epithelial sodium channel cause salt wasting with hyperkalaemic acidosis, pseudohypoaldosteronism type 1. *Nat. Genet* **12**, 248-253 (1996).
78. S. S. Strautnieks, R. J. Thompson, R. M. Gardiner, E. Chung, A novel splice-site mutation in the gamma subunit of the epithelial sodium channel gene in three pseudohypoaldosteronism type 1 families. *Nat. Genet* **13**, 248-250 (1996).
79. Y. S. Oh, D. G. Warnock, Disorders of the epithelial Na+ channel in Liddle's syndrome and autosomal recessive pseudohypoaldosteronism type 1. *Nephron Experimental Nephrology* **8**, 320-325 (2000).
80. R. A. Shimkets *et al.*, Liddle's syndrome: Heritable human hypertension caused by mutations in the b subunit of the epithelial sodium channel. *Cell* **79**, 407-414 (1994).
81. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
82. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359 (2012).
83. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
84. J. O'Rawe *et al.*, Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome med* **5**, 28 (2013).
85. K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
86. E. Muhammad *et al.*, Autosomal recessive hyponatremia due to isolated salt wasting in sweat associated with a mutation in the active site of Carbonic Anhydrase 12. *Hum. Genet* **129**, 397-405 (2011).
87. M. Feldshtein *et al.*, Hyperchlorhidrosis caused by homozygous mutation in CA12, encoding carbonic anhydrase XII. *Am J Hum. Genet* **87**, 713-720 (2010).

88. Y. Feinstein *et al.*, Natural history and clinical manifestations of hyponatremia and hyperchlorhidrosis due to carbonic anhydrase XII deficiency. *Hormone Research in Paediatrics* **81**, 336-342 (2014).
89. F. G. Riepe *et al.*, Revealing a subclinical salt-losing phenotype in heterozygous carriers of the novel S562P mutation in the α subunit of the epithelial sodium channel. *Clinical endocrinology* **70**, 252-258 (2009).
90. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*, (2015).
91. K. Mosler *et al.*, Feasibility of nasal epithelial brushing for the study of airway epithelial functions in CF infants. *Journal of Cystic Fibrosis* **7**, 44-53 (2008).
92. J. Haapasalo *et al.*, Identification of an alternatively spliced isoform of carbonic anhydrase XII in diffusely infiltrating astrocytic gliomas. *Neuro-oncology* **10**, 131-138 (2008).
93. M. Uhlen *et al.*, Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
94. L. E. Maquat, Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol* **5**, 89-99 (2004).
95. D. A. Whittington *et al.*, Crystal structure of the dimeric extracellular domain of human carbonic anhydrase XII, a bitopic membrane protein overexpressed in certain cancer tumor cells. *Proc Natl Acad Sci U S A* **98**, 9545-9550 (2001).
96. M. Mall *et al.*, Effect of genistein on native epithelial tissue from normal individuals and CF patients and on ion channels expressed in Xenopus oocytes. *British journal of pharmacology* **130**, 1884-1892 (2000).
97. S. Parkkila *et al.*, Expression of the membrane-associated carbonic anhydrase isozyme XII in the human kidney and renal tumors. *Journal of Histochemistry & Cytochemistry* **48**, 1601-1608 (2000).
98. M. Lee, *Basic skills in interpreting laboratory data*. (ASHP, 2009).
99. P. M. Quinton, Cystic fibrosis: impaired bicarbonate secretion and mucoviscidosis. *The Lancet* **372**, 415-417 (2008).
100. C. R. Scriver, S. Kaufman, in *Metabolic and Molecular Bases of Inherited Disease,* C. R. Scriver, A. L. Beaudet, D. Valle, W. S. Sly, Eds. (McGraw-Hill, Inc., New York, 2001), chap. 77, pp. 1667-1724.
101. W. S. Sly, G. Shah, in *The Metabolic and Molecular Bases of Inherited Disease,* C. R. Scriver, A. L. Beaudet, W. S. Sly, D. Valle, Eds. (McGraw-Hill, Inc., New York, 2001), vol. IV, chap. 208, pp. 5331-5343.
102. K. Gilmour, Perspectives on carbonic anhydrase. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* **157**, 193-197 (2010).
103. R. E. Tashian, D. HEWETT-EMMETT, S. J. Dodgson, R. E. Forster, W. S. Sly, The Value of Inherited Deficiencies of Human Carbonic Anhydrase Isozymes in Understanding Their Cellular Rolesa. *Annals of the New York Academy of Sciences* **429**, 262-275 (1984).
104. A. Kivelä *et al.*, Expression of a novel transmembrane carbonic anhydrase isozyme XII in normal human gut and colorectal tumors. *The American journal of pathology* **156**, 577-584 (2000).
105. P. Karhumaa *et al.*, Identification of carbonic anhydrase XII as the membrane isozyme expressed in the normal human endometrial epithelium. *Molecular human reproduction* **6**, 68-74 (2000).
106. H. McMurtrie *et al.*, The bicarbonate transport metabolon. *Journal of enzyme inhibition and medicinal chemistry* **19**, 231-236 (2004).
107. D. Sterling, B. V. Alvarez, J. R. Casey, The Extracellular Component of a Transport Metabolon EXTRACELLULAR LOOP 4 OF THE HUMAN AE1 Cl−/HCO

EXCHANGER BINDS CARBONIC ANHYDRASE IV. *Journal of Biological Chemistry* **277**, 25239-25246 (2002).
108. P. E. Morgan, S. Pastorekova, A. K. Stuart-Tilley, S. L. Alper, J. R. Casey, Interactions of transmembrane carbonic anhydrase, CAIX, with bicarbonate transporters. *Am. J. Physiol Cell Physiol* **293**, C738-C748 (2007).
109. J. R. Casey, W. S. Sly, G. N. Shah, B. V. Alvarez, Bicarbonate homeostasis in excitable tissues: role of AE3 Cl−/HCO3− exchanger and carbonic anhydrase XIV interaction. *American Journal of Physiology-Cell Physiology* **297**, 1091-1102 (2009).
110. L. Guglani, B. Sitwat, D. Lower, G. Kurland, D. J. Weiner, Elevated sweat chloride concentration in children without cystic fibrosis who are receiving topiramate therapy. *Pediatr. Pulmonol* **47**, 429-433 (2012).
111. J. Y. Winum, S. A. Poulsen, C. T. Supuran, Therapeutic applications of glycosidic carbonic anhydrase inhibitors. *Medicinal research reviews* **29**, 419-435 (2009).
112. C. T. Supuran, Carbonic anhydrases: novel therapeutic applications for inhibitors and activators. *Nature reviews Drug discovery* **7**, 168-181 (2008).
113. N. Mirza, A. G. Marson, M. Pirmohamed, Effect of topiramate on acid–base balance: extent, mechanism and effects. *British journal of clinical pharmacology* **68**, 655-661 (2009).
114. D. Weycker, J. Edelsberg, G. Oster, G. Tino, Prevalence and economic burden of bronchiectasis. *Clinical Pulmonary Medicine* **12**, 205-209 (2005).
115. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
116. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
117. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010).
118. M. A. Depristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet* **43**, 491-498 (2011).
119. S. B. Ng *et al.*, Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-276 (2009).
120. M. Yandell *et al.*, A probabilistic disease-gene finder for personal genomes. *Genome Res*, (2011).
121. X. Li *et al.*, Integrin alpha6beta4 identifies human distal lung epithelial progenitor cells with potential as a cell-based therapy for cystic fibrosis lung disease. *PLoS One* **8**, e83624 (2013).
122. X. Li *et al.*, CFTR is required for maximal transepithelial liquid transport in pig alveolar epithelia. *American journal of physiology. Lung cellular and molecular physiology* **303**, L152-160 (2012).
123. G. L. Peterson, Review of the Folin phenol protein quantitation method of Lowry, Rosebrough, Farr and Randall. *Analytical biochemistry* **100**, 201-220 (1979).
124. T. H. Maren, A simplified micromethod for the determination of carbonic anhydrase and its inhibitors. *Journal of Pharmacology and Experimental Therapeutics* **130**, 26-29 (1960).
125. V. Sundaram, P. Rumbolo, J. Grubb, P. Strisciuglio, W. S. Sly, Carbonic anhydrase II deficiency: diagnosis and carrier detection using differential enzyme inhibition and inactivation. *American journal of human genetics* **38**, 125 (1986).
126. G. Veit *et al.*, From CFTR biology toward combinatorial pharmacotherapy: expanded classification of cystic fibrosis mutations. *Mol Biol Cell* **27**, 424-433 (2016).
127. N. Morral *et al.*, The origin of the major cystic fibrosis mutation (deltaF508) in European populations. *Nature Genet* **7**, 169-175 (1994).

128. J. L. Bobadilla, M. Macek, J. P. Fine, P. M. Farrell, Cystic fibrosis: A worldwide analysis of CFTR mutations - Correlation with incidence data and application to screening. *Human Mutation* **19**, 575-606 (2002).
129. B. H. Qu, E. Strickland, P. J. Thomas, Cystic fibrosis: a disease of altered protein folding. *J Bioenerg. Biomembr* **29**, 483-490 (1997).
130. D. C. Gadsby, P. Vergani, L. Csanady, The ABC protein turned chloride channel whose failure causes cystic fibrosis. *Nature* **440**, 477-483 (2006).
131. E. Kerem, J. Reisman, M. Corey, G. J. Canny, H. Levison, Prediction of mortality in patients with cystic fibrosis. *N. Engl. J Med* **326**, 1187-1191 (1992).
132. C. E. Wainwright *et al.*, Lumacaftor-Ivacaftor in Patients with Cystic Fibrosis Homozygous for Phe508del CFTR. *N. Engl. J. Med*, (2015).
133. F. J. Accurso *et al.*, Effect of VX-770 in persons with cystic fibrosis and the G551D-CFTR mutation. *N. Engl. J Med* **363**, 1991-2003 (2010).
134. G. L. Lukacs, A. S. Verkman, CFTR: folding, misfolding and correcting the DeltaF508 conformational defect. *Trends Mol. Med* **18**, 81-91 (2012).
135. C. L. Ward, S. Omura, R. R. Kopito, Degradation of CFTR by the ubiquitin-proteasome pathway. *Cell* **83**, 121-127 (1995).
136. T. J. Jensen *et al.*, Multiple proteolytic systems, including the proteasome, contribute to CFTR processing. *Cell* **83**, 129-135 (1995).
137. B. H. Qu, P. J. Thomas, Alteration of the cystic fibrosis transmembrane conductance regulator folding pathway. *J. Biol. Chem* **271**, 7261-7264 (1996).
138. R. R. Kopito, Biosynthesis and degradation of CFTR. *Physiol Rev* **79**, S167-S173 (1999).
139. W. Dalemans *et al.*, Altered chloride ion channel kinetics associated with the deltaF508 cystic fibrosis mutation. *Nature* **354**, 526-528 (1991).
140. Z. Kopeikin, Z. Yuksek, H. Y. Yang, S. G. Bompadre, Combined effects of VX-770 and VX-809 on several functional abnormalities of F508del-CFTR channels. *J. Cyst. Fibros*, (2014).
141. F. Van Goor *et al.*, Correction of the F508del-CFTR protein processing defect in vitro by the investigational drug VX-809. *Proc. Natl. Acad. Sci. U. S. A* **108**, 18843-18848 (2011).
142. T. Okiyoneda *et al.*, Mechanism-based corrector combination restores DeltaF508-CFTR folding and function. *Nat. Chem. Biol* **9**, 444-454 (2013).
143. P. W. Phuan *et al.*, Synergy-based Small-Molecule Screen Using a Human Lung Epithelial Cell Line Yields DeltaF508-CFTR Correctors that Augment VX-809 Maximal Efficacy. *Mol. Pharmacol* **86**, 42-51 (2014).
144. S. Kiesewetter *et al.*, The CFTR mutation R117H produces different phenotypes depending on genetic background. *Am. J. Med. Genet* **53**, #86 (1993).
145. B. Tummler, F. Stanke, I. Bronsveld, H. Veeze, M. Ballmann, Transient correction of the basic defect in sweat glands in an individual with cystic fibrosis carrying the complex CFTR allele F508del-R553Q. *Thorax* **64**, 179-180 (2009).
146. A. Hamosh, M. Corey, Correlation between genotype and phenotype in patients with cystic fibrosis. The Cystic Fibrosis Genotype-Phenotype Consortium. *N. Engl. J Med* **329**, 1308-1313 (1993).
147. N. P. Blackledge, C. J. Ott, A. E. Gillen, A. Harris, An insulator element 3' to the CFTR gene binds CTCF and reveals an active chromatin hub in primary cells. *Nucleic Acids Res* **37**, 1086-1094 (2009).
148. R. Yang *et al.*, Differential contribution of cis-regulatory elements to higher order chromatin structure and expression of the CFTR locus. *Nucleic Acids Res*, (2015).
149. A. Sobczyńska-Tomaszewska *et al.*, Newborn screening for cystic fibrosis: Polish 4 years' experience with CFTR sequencing strategy. *Eur J Hum Genet* **21**, 391-396 (2013).

150. P. Kolesár *et al.*, Mutation analysis of the CFTR gene in Slovak cystic fibrosis patients by DHPLC and subsequent sequencing: identification of four novel mutations. *Gen Physiol Biophys* **27**, 299-305 (2008).
151. F. Amato *et al.*, Extensive molecular analysis of patients bearing CFTR-related disorders. *J Mol Diagn* **14**, 81-89 (2012).
152. E. Elahi *et al.*, A haplotype framework for cystic fibrosis mutations in Iran. *J Mol Diagn* **8**, 119-127 (2006).
153. L. S. Smit, D. J. Wilkinson, M. K. Mansoura, F. S. Collins, D. C. Dawson, Functional roles of the nucleotide-binding folds in the activation of the cystic fibrosis transmembrane conductance regulator. *Proc. Natl. Acad. Sci. U. S. A* **90**, 9963-9967 (1993).
154. A. El-Seedy *et al.*, Influence of the duplication of CFTR exon 9 and its flanking sequences on diagnosis of cystic fibrosis mutations. *J Mol Diagn* **11**, 488-493 (2009).
155. R. Rozmahel *et al.*, Amplification of CFTR exon 9 sequences to multiple locations in the human genome. *Genomics* **45**, 554-561 (1997).
156. N. Gheldof *et al.*, Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. *Nucleic Acids Res* **38**, 4325-4336 (2010).
157. E. M. Smith, B. R. Lajoie, G. Jain, J. Dekker, Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus. *Am J Hum Genet* **98**, 185-201 (2016).
158. N. Gheldof *et al.*, Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. *Nucleic Acids Res* **38**, 4325-4336 (2010).
159. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol* **9**, e1001046 (2011).
160. T. E. Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640 (2004).
161. O. Delaneau, J. F. Zagury, J. Marchini, Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6 (2013).
162. S. Cordovado *et al.*, CFTR mutation analysis and haplotype associations in CF patients. *Molecular genetics and metabolism* **105**, 249-254 (2012).
163. Y. Fichou *et al.*, Estimating the age of CFTR mutations predominantly found in Brittany (Western France). *J. Cyst. Fibros* **7**, 168-173 (2008).
164. J. M. Rommens *et al.*, Identification and Regional Localization of DNA Markers on Chromosome 7 for the Cloning of the Cystic Fibrosis Gene. *Am. J. Hum. Genet* **43**, 645-663 (1988).
165. V. A. McCarthy, A. Harris, The CFTR gene and regulation of its expression. *Pediatr. Pulmonol* **40**, 1-8 (2005).
166. C. J. Ott *et al.*, A complex intronic enhancer regulates expression of the CFTR gene by direct interaction with the promoter. *J Cell Mol. Med* **13**, 680-692 (2009).
167. F. C. Broackes-Carter *et al.*, Temporal regulation of CFTR expression during ovine lung development: implications for CF gene therapy. *Human molecular genetics* **11**, 125-131 (2002).
168. N. Gosalia, A. Harris, Chromatin dynamics in the regulation of CFTR expression. *Genes* **6**, 543-558 (2015).
169. S. Moisan *et al.*, Analysis of long-range interactions in primary human cells identifies cooperative CFTR regulatory elements. *Nucleic acids research*, gkv1300 (2015).
170. F. Stanke *et al.*, The CF-modifying gene EHF promotes p. Phe508del-CFTR residual function by altering protein glycosylation and trafficking in epithelial cells. *European Journal of Human Genetics* **22**, 660-666 (2014).
171. H. J. Veeze, J. J. Halley, J. C. deJongste, H. R. deJonge, M. Sinaasappel, Determinants of mild clinical symptoms in cystic fibrosis patients. *J. Clin. Invest* **93**, 461-466 (1994).

172. D. Penque *et al.*, Cystic fibrosis F508del patients have apically localized CFTR in a reduced number of airway cells. *Laboratory investigation* **80**, 857-868 (2000).
173. J. F. Dekkers, C. K. van der Ent, J. M. Beekman, Novel opportunities for CFTR-targeting drug development using organoids. *Rare Diseases* **1**, 939-945 (2013).
174. A. Van Barneveld *et al.*, Functional analysis of F508del CFTR in native human colon. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **1802**, 1062-1069 (2010).
175. T. Boulikas, Chromatin domains and prediction of MAR sequences. *International review of cytology* **162**, 279-388 (1996).
176. G. Devi, Y. Zhou, Z. Zhong, D. F. K. Toh, G. Chen, RNA triplexes: from structural principles to biological and biotech applications. *Wiley Interdisciplinary Reviews: RNA* **6**, 111-128 (2015).
177. M. Gymrek *et al.*, Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet* **48**, 22-29 (2016).
178. K. K. Deeb *et al.*, The c.1364C>A (p.A455E) Mutation in the CFTR Pseudogene Results in an Incorrectly Assigned Carrier Status by a Commonly Used Screening Platform. *J Mol Diagn* **17**, 360-365 (2015).
179. T. I. H. Consortium, The International HapMap Project. *Nature* **426**, 789-796 (2003).
180. G. Zhang, D. W. Nebert, R. Chakraborty, L. Jin, Statistical power of association using the extreme discordant phenotype design. *Pharmacogenet. Genomics* **16**, 401-413 (2006).
181. . (2009).
182. C. Taylor *et al.*, A novel lung disease phenotype adjusted for mortality attrition for cystic fibrosis genetic modifier studies. *Pediatric Pulmonology* **Epub no. doi: 10.1002/ppul.21456**, (2011).
183. P. a. C. C. M. University of North Carolina at Chapel Hill.
184. S. Purcell *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum. Genet* **81**, 559-575 (2007).
185. S. Lee, M. C. Wu, X. Lin, Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762-775 (2012).
186. E. Minikel, in *CureFFI.org*. (2012).
187. J. H. Hong *et al.*, Essential role of carbonic anhydrase XII in secretory gland fluid and HCO3− secretion revealed by disease causing human mutation. *The Journal of physiology* **593**, 5299-5312 (2015).
188. C. K. King, R. Glass, J. S. Bresee, C. Duggan, Managing acute gastroenteritis among children. *MMWR Recomm Rep* **52**, 1-16 (2003).
189. R. Maughan, Impact of mild dehydration on wellness and on exercise performance. *European Journal of Clinical Nutrition* **57**, S19-S23 (2003).
190. M. Danz, K. Pöttgen, P. M. Tönjes, J. Hinkelbein, S. Braunecker, Hyponatremia among Triathletes in the Ironman European Championship. *New England Journal of Medicine* **374**, 997-998 (2016).
191. J. K. Gustafsson *et al.*, Bicarbonate and functional CFTR channel are required for proper mucin secretion and link cystic fibrosis with its mucus phenotype. *The Journal of experimental medicine* **209**, 1263-1272 (2012).
192. P. Karp *et al.*, Airway Surface Liquid (asl) Ph And Bicarbonate Secretion In Cultured Small Airway Cells From Cystic Fibrosis Pigs. *Am J Respir Crit Care Med* **191**, A5999 (2015).
193. S. Tate, G. MacGregor, M. Davis, J. A. Innes, A. P. Greening, Airways in cystic fibrosis are acidified: detection by exhaled breath condensate. *Thorax* **57**, 926-929 (2002).
194. V. S. Shah *et al.*, Airway acidification initiates host defense abnormalities in cystic fibrosis mice. *Science* **351**, 503-507 (2016).

195. Y. Yang *et al.*, Molecular findings among patients referred for clinical whole-exome sequencing. *Jama* **312**, 1870-1879 (2014).
196. D. Lvovs, O. Favorova, A. Favorov, A polygenic approach to the study of polygenic diseases. *Acta Naturae (англоязычная версия)* **4**, (2012).
197. N. R. Wray *et al.*, Research review: polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry* **55**, 1068-1087 (2014).
198. J. Lieberman, Cyclic fluctuation of sweat electrolytes in women: Effect of polythiazide upon sweat electrolytes. *JAMA* **195**, 629-635 (1966).

CURRICULUM VITAE

Briana Vecchio-Pagán

Education

Johns Hopkins - School of Medicine
Doctorate of Philosophy, Cellular and Molecular Medicine
Laboratory of Dr. Garry R. Cutting – McKusick-Nathans Institute of Genetic Medicine

West Virginia University - Honors College
Bachelor of Arts, Chemistry (2010)
Bachelor of Arts, Biochemistry (2010)
Summa Cum Laude

Technical Competence

Operating Systems: UNIX (Ubuntu, Redhat, Fedora), Windows
Coding Languages: Perl (Advanced), Java (Beginner), Python(Beginner), Bash/Shell(Beginner)
Statistical Packages: R (+ Bioconductor), STATA, SAS
Database Management Systems: MySQL, SQL, ACCESS

Genetic Analyses

Next Generation Sequencing (NGS) Analysis Pipelines (Whole Genome, Exome, & RNA-seq)
Genome Wide Association Studies (GWAS) of both microarray genotyped and imputed data
Genetic Linkage and Familial TDT

Publications

Deep resequencing of *CFTR* in 762 F508del homozygotes reveals clusters of non-coding variants associated with variation in sweat chloride concentration and lung function. **Briana Vecchio-Pagán**, S Blackman, KS Raraigh, RG Pace, MJ Pellicore, A Franca, M Lee, N Sharma, MR Knowles, GR Cutting. *Submitted to Human Genetics*.

Codon Bias and the Folding Dynamics of the Cystic Fibrosis Transmembrane Conductance Regulator. R Bartoszewski, J Króliczewski, A Piotrowski, A Janaszak-Jasiecka, S Bartoszewska, **Briana Vecchio-Pagán**, L Fu, S Matalon, GR Cutting, SM Rowe, and JF Collawn. *Submitted to PLOS ONE*.

Loss of carbonic anhydrase XII function in individuals with elevated sweat chloride concentration and pulmonary airway disease. **Briana Vecchio-Pagán** , M Lee, N Sharma, A Waheed, X Li, K Raraigh, S Robbins, S Han, A Franca, M Pellicore, T Evans, H Nguyen, S Luan, D Belchis, J Hertecant, J Zabner, W Sly, GR Cutting. *Human Molecular Genetics*, 2016.

Creation and characterization of an airway epithelial cell line for stable expression of CFTR variants. L Gottschalk, **Briana Vecchio-Pagán**, N Sharmaa, S Han, A Franca, E Wohlerb, D Batista, GR Cutting. *Cystic Fibrosis*, 2015.

Experimental assessment of splicing variants using expression minigenes and comparison with in silico predictions. N Sharma, PR Sosnay, AS Ramalho, C Douville, A Franca, L Gottschalk, J Park,

M Lee, **Briana Vecchio-Pagán**, K Raraigh, M Amaral, R Karchin, GR Cutting. *Human Mutation*, 2014.

Cigarette smoke induces epithelial to mesenchymal transition and increases the metastatic ability of breast cancer cells. V Francescopaolo Di Cello, H Li**, B Vecchio-Pagán**, B Gordon, K Harbom, James Shin, Robert Beaty, Wei Wang, Cory Brayton, Stephen B Baylin, Cynthia A Zahnow. *Molecular Cancer*, 2013.

Undergraduate Awards / Distinctions

- WVU Foundation Outstanding Senior (1 of 35 selected in the class of ~5,000)
- 2009 Undergraduate Researcher of the Year STaR Symposium (1/~500 abstracts & 20 presentations)
- WV Nano Symposium (2010, 2nd place, 2009, 3rd place / ~200 select participants)
- Merk Index Award – Awarded to A Senior Expected to Excel in a Biomedical Career

Oral Presentations

2015 29th Annual North American Cystic Fibrosis Conference
Common and Rare Variation Associated with the F508del Mutation Identified by Deep Re-sequencing of CFTR in 602 Cystic Fibrosis Patients

2015 8th Annual JHU-UMD Diabetes Research Center Symposium
Cellular Senescence in Cystic Fibrosis Related Diabetes: The role of telomeres and the CDKN2AB locus

2014 International Cystic Fibrosis Genetic Modifiers Consortium Face to Face Meeting
SLC26A9: Associations with CFRD, MI, SAKNORM, & functional assays of 5' promotor variants

2012 Johns Hopkins Cystic Fibrosis Research Seminar
Beyond CFTR: Exome Sequencing to Identify Novel Genes Causing Atypical Cystic Fibrosis Phenotypes

Poster Presentations / Grants

2014 North American Cystic Fibrosis Conference
Elevated Sweat Chloride Levels in Atypical CF Patients Indicate Additional Genotyping of CA XII

2014 American Society of Human Genetics National Meeting
Deep re-sequencing of CFTR bearing the common F508del mutation reveals a rare variant associating with variation in lung infection

2013 American Society of Human Genetics National Meeting
Loss of function mutations in Carbonic Anhydrase XII result in hyponatremic dehydration and elevated sweat chloride concentration.

2010 American Chemistry Society National Meeting
Evolution of single-stranded DNA molecular recognition elements via CE-SELEX: Detection of TNT and biosensor applications

Grant Award: Complete Genomics Whole Genome Sequencing Initiative
VAAST: A novel variant annotation and prioritization tool for determining the causative alleles of atypical Cystic Fibrosis. Co-author and awardee of ~$80,000.

Previous Research Experiences

2009 Summer Intern: University of Pennsylvania and CHoP Gene Therapy Program

-The role of VAMP-3 in lamellar body formation, trafficking, and fusion pathways during fetal lun development.

2008 Summer Intern: University of Pittsburgh School of Medicine Department of Neuropharmacology

-The role of select phosphatases in cell signaling pathways crucial to mitochondrial motility in primary cortical neurons.

2008-10 Undergraduate Research: West Virginia University Eberly College – Sooter Lab

-Molecular Recognition Elements (MREs) generated via in vitro selection or SELEX of an ssDNA library to detect TNT – for use in detection of improvised explosive devices.

Professional Affiliations

- American Chemistry Society (2006-2010)
- American Society of Human Genetics (2012-Current)

Teaching Experience

2007-2010 Tutoring Chemistry, Biology, and Mathematics – West Virginia University

2008-2010 Peer Led Team Learning – West Virginia University

2009-2010 General Chemistry II Teaching Assistant – West Virginia University

2010-2014 Private Tutoring High School Math and Sciences